Unified Framework of Feature Based Adaptation for Statistical Speech Synthesis and Recognition

THÈSE N^{O} 5612 (2013)

PRÉSENTÉE LE 4 AVRIL 2013

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

LABORATOIRE DE L'IDIAP

PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Lakshmi BABU SAHEER

acceptée sur proposition du jury:

Dr J.-M. Vesin, président du jury
Prof. H. Bourlard, directeur de thèse
A. W. Black, rapporteur
Dr R. Schlüter, rapporteur
Prof. J.-Ph. Thiran, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2013

Faith is taking the first step even when
you don't see the whole staircase.
— Martin Luther King, Jr.

To my loving husband...

Acknowledgements

These past four years has been a wonderful experience for me. It has been a tough and equally exciting journey. I could not have completed it successfully without the love and support from a lot of people around me. Firstly, I thank my supervisors Philip N. Garner and John Dines for their dedicated support and guidance. Even though EPFL declined to recognise them officially as my supervisors, they warmly extended their generous help at every step of my PhD. This research would not have been fruitful without their encouragement and guidance. They gave great advices and always showed me the right direction. I really envy the great technical clarity that Phil has for any topic we discussed and I learned a lot from him than from any textbooks or research papers. Phil would be always around (anytime of the day) and patiently correct every bit of my technical writing usually in the late evenings that too in pubs. John on the other hand was a great source of motivation, supporting every activity that I wanted to pursue (even if they sounded stupid). He always managed to find time for reviewing my work in spite of his busy schedule as a start-up CTO. I consider myself extremely lucky to have you both as my supervisors.

I would also like to thank Dr. Junichi Yamagishi from the centre for speech technology research (CSTR), Edinburgh for all his guidance and collaboration. He helped me with some important parts of my PhD research and provided most of the data and scripts for my work. This collaboration was very fruitful as we published a conference paper and are in the process of writing a journal paper. He always came up with a number of ideas for research and even helped me personally in my trips to Japan for conferences. He acted like a co-supervisor to me with his guidance and performed some subjective listening tests for me at CSTR. It was a pleasure to work with my PhD colleague and friend, Hui Liang. We were both working for the Effective multilingual interaction in mobile environments (EMIME) project and helped and supported each other with healthy discussions and debates. I was happy to have him around to share the work and ideas. We did a lot of research together mainly during the beginning of the project and the cross-lingual work presented in this thesis. He was a great companion to have especially during the travels for conferences and project meetings.

It was a great opportunity to work on the EMIME project with the different partners. It was a pleasure to meet all the partners at the project meetings and discuss different technical aspects of the project apart from having fun. I thank Prof. Steve Renals, Dr. Simon King (both from CSTR and University of Edinburgh), Dr. Mikko Kurimo (Aalto University, Finland), Prof. Keiichi Tokuda, (Nagoya Institute of Technology, Japan), Dr. Jilei Tian (Nokia, China) and many others. I extend my gratitude to my funding sources, the EMIME (European Union FP7) project

and V-FAST (Hasler) project for the financial support in the course of my doctoral studies.
I would like to thank Idiap management especially Prof. Hervé Bourlard (my supervisor and director of Idiap) for providing me this great opportunity to work at Idiap and ensuring all the resources needed for my research. Prof. Bourlard was a great source of personal inspiration. I thank the secretaries, Mrs. Nadine Rousseau and Mrs. Sylvie Millius for all the administrative support. It was not easy to find my way out in Switzerland from the first day of my arrival till date. Special thanks to the deputy director of Idiap, Dr. Francois Foglia for his support specially during the international create challenge (ICC 2012) and his support and confidence in my project. I thank Dr. Milos Cernak for the all the help and support especially for the ICC project. We are a great team and plan to continue this collaboration as far as we can. The ICC group was a good source of happiness. I also thank the other support staff at Idiap, Frank Formaz, Norbert Crettol, Vincent Spano, Alexandre Nanchen, Ed Gregg, Christophe Ecoeur and several others. Special thanks to my colleagues Flavio Tarsetti and Laurent El-Shafey for the help with French translations. I am lucky to have known Particia Emonet with her help in French and my personal life.
Special thanks to my friend and colleague, Afsaneh Asaei for being there to share my happiness and to comfort me in times of distress. We helped each other in our ordeals. Similarily, my friend and colleague, Ramya Rasipuram for being a great source of comfort. We both delivered our babies around the same time and could easily share our happiness and troubles. We had some great hikes organised by Marco Fornoni, Laurent El-Shafey, and Deepu Vijayasenan. Thanks to the Indian community in Martigny (Samuel, Jamie, Deepu, Ramya, Murali, Venkatesh, Abhilasha, Jagan, Gokul, Sriram, Harsha, Mathew, Dinesh and several others) for keeping my social life alive with great activities like Indian dinner night, barbecues and other get-togethers from time to time. There a lot of other Idiap colleagues who made my PhD life enjoyable including Oya, Daiira, Joan, Paco, Serena, Tatiana, Niklas, Gelareh, Elham, Sameera, Jean-marc, Petr, Nicolae, and Aileen to name a few.
I extend my warm gratitude to all my family and friends. Specially, my father, mother and sister for their encouragement and love. Also, my father-in-law and mother-in-law for their love and support. Finally, I am extremely happy to have a wonderful husband, Saheer and an adorable daughter, Nora without whose unconditional love and support, this PhD would have been impossible. Although she is just one year old, my daughter co-operated with every single activity I had to take up. My husband is always a source of encouragement, convincing me to take up impossible tasks, boosting my confidence by enlightening me with my potential and helping me in every way he can. I am really fortunate to have him as my life partner. I dedicate this work to him.
The list of people mentioned in this acknowledgement is not exhaustive and I apologise in case I accidentally missed out some of the very special people who have influenced my PhD life.
Martigny, 28 November 2012 Lakshmi Saheer.

Abstract

The advent of statistical parametric speech synthesis has paved new ways to a unified framework for hidden Markov model (HMM) based text to speech synthesis (TTS) and automatic speech recognition (ASR). The techniques and advancements made in the field of ASR can now be adopted in the domain of synthesis. Speaker adaptation is a well-advanced topic in the area of ASR, where the adaptation data from a target speaker is used to transform the canonical model parameters to represent a speaker specific model. Feature adaptation techniques like vocal tract length normalization (VTLN) perform the same task by transforming the features ; this can be shown to be equivalent to model transformation. The main advantage of VTLN is that it can demonstrate noticeable improvements in performance with very little adaptation data and can be classified as a rapid adaptation technique.

VTLN is a widely used technique in ASR, and can be used in TTS to improve the rapid adaptation performance. In TTS, the task is to synthesize speech that sounds like a particular target speaker. Using VTLN for TTS is found to make the output synthesized speech sound quite similar to the target speaker from his very first utterance. An all-pass filter based bilinear transform was implemented for the mel-generalized cepstral (MGCEP) features of the HMM-based speech synthesis system (HTS). The initial implementation was using a grid search approach that selects the best warping factor for the speech spectrum from a grid of available values using the maximum likelihood criterion. VTLN was shown to give performance improvements in the rapid adaptation framework where the number of adaptation sentences from the target speaker was limited. But, this technique involves huge time and space complexities and the rapid adaptation demands for an efficient implementation of the VTLN technique.

To this end, an efficient expectation maximization (EM) based VTLN approach was implemented for HTS using Brent's search. Unlike the ASR features, MGCEP does not use a filter bank (in order to facilitate the speech reconstruction) and this provides equivalence to the model transformation for the EM implementation. This facilitates the estimation of warping factors to be embedded in the HMM training using the same sufficient statistics as in constrained maximum likelihood linear regression (CMLLR). This work addresses a lot of challenges faced in the process of adopting VTLN for synthesis due to the higher dimensionality of the cepstral features used in the TTS models. The main idea was to unify the theory and practise in the implementation of VTLN for both ASR and TTS. Several techniques have been proposed in this thesis, in order to find the best feasible warping factor estimation procedure. Estimation of the warping factor using the lower order cepstral features representing the spectral envelope is demonstrated to be the best approach. Different evaluations on standard databases are

performed in this work to illustrate the performance improvements and perceptual challenges involved in the VTLN adaptation.
VTLN has only a single parameter to represent the speaker characteristics and hence, has the limitation of not scaling to the performance of other linear transform based adaptation methods with the availability of large amounts of adaptation data. Several techniques are demonstrated in this work to combine the model based adaptation like constrained structural maximum a posteriori linear regression (CSMAPLR) with VTLN, one such technique being using VTLN as the prior transform at the root node of the tree structure of the CSMAPLR system. Thus, along with rapid adaptation, the performance scales with the availability of more adaptation data. These techniques although developed for TTS, can also be effectively used in ASR. It was also shown to give improvements in ASR especially for scenarios like noisy speech conditions. Other improvements to rapid adaptation including a bias term for VTLN, multiple transform based VTLN using regression classes and VTLN prior for non-structural MAPLR adaptation are also proposed. These techniques also demonstrated both ASR and TTS performance improvements. Also, a few special scenarios, specifically cross-lingual speech, cross-gender speech, child speech and noisy speech evaluations are presented where the rapid adaptation methods presented in this work was shown to be highly beneficial. Most of these methods will be published as extensions to the open-source HTS toolkit.

Keywords Vocal tract length normalization, Mel-generalized cepstral features, All-pass filter based bilinear transformations, Rapid feature adaptation, HMM-based statistical parametric speech synthesis (HTS), HMM-based automatic speech recognition (ASR), Unified modeling and adaptation of ASR and TTS, Expectation maximization, Model transformations, Constrained structural maximum a posteriori linear regression, Constrained likelihood linear regression.

Résumé

L'introduction de méthodes statistiques paramétriques pour la synthèse de la parole a ouvert la voie à un cadre unifié pour la synthèse de parole à partir d'une entrée textuelle (SPET) et la reconnaissance automatique de la parole (RAP) reposant sur des modèles de Markov cachés (MCC). Les techniques et les améliorations effectuées pour la RAP peuvent désormais être adoptées dans le domaine de la synthèse. L'adaptation au locuteur est un sujet très développé dans le domaine de la RAP, où des données d'adaptation issues d'un locuteur cible sont utilisées pour modifier les paramètres d'un modèle générique initial afin d'obtenir un modèle spécifique au locuteur. Des techniques d'adaptation de primitives comme la normalisation de la longueur du conduit vocal (NLCV) effectuent la même tâche en modifiant les primitives ; il peut être établi que cette approche est équivalente à une transformation du modèle. Le principal avantage de la NLCV est que des améliorations significatives des performances ont été constatées lorsque peu de données d'adaptation sont disponibles, et qu'elle peut être considérée comme une technique d'adaptation rapide.

La NLCV est une technique très répandue pour la RAP, et elle peut être utilisée pour la SPET afin de permettre une adaptation rapide. Pour la SPET, l'objectif est de synthétiser la parole de sorte qu'elle ressemble à celle d'un locuteur cible donné. Il a été constaté que l'utilisation de la NLCV pour la SPET conduit à de la parole synthétisée qui ressemble à celle du locuteur cible, et ce, dès le début de l'élocution. Une transformation bilinéaire reposant sur un filtre déphaseur a été implémentée pour les coefficients cepstraux sur l'échelle de Mel généralisée (CCEMG) utilisés par le système de synthèse de la parole basé sur des MCC (SSPM). L'implémentation initiale utilise une méthode de recherche par quadrillage qui sélectionne le facteur de déformation optimal pour le spectre vocal à partir d'une grille de valeurs disponibles en utilisant le critère de maximum de vraisemblance. Il a été établi que la NLCV conduit à une amélioration des performances dans un cadre d'adaptation rapide où le nombre de phrases pour l'adaptation au locuteur cible est limité. Toutefois, cette technique a une complexité en temps et en espace mémoire importante, alors que l'adaptation rapide exige une implémentation efficace de la technique NLCV.

A cette fin, une approche efficace de la NLCV reposant sur un algorithme d'espérance-maximisation (EM) a été implémentée pour le SSPM en utilisant la méthode de Brent. Contrairement aux primitives utilisées pour la RAP, les CCEMG n'emploient pas de batterie de filtres (afin de faciliter la reconstruction de la parole), ce qui les rapproche de la transformation de modèle pour l'implémentation EM. Cela simplifie l'estimation des facteurs de déformation qui peut être intégrée dans la phase d'entraînement des MCC en utilisant les mêmes

statistiques exhaustives que pour la régression linéaire par maximum de vraisemblance avec contraintes (RLMVC). Cette thèse aborde les nombreux défis rencontrés en vue d'utiliser la NLCV pour la synthèse, principalement en raison de la dimension plus grande des primitives cepstrales utilisées pour les modèles de la SPET. L'idée centrale est d'unifier la théorie et la pratique dans l'implémentation de la NLCV pour à la fois la RAP et la SPET. De nombreuses techniques sont proposées dans cette thèse, afin de trouver la meilleure procédure réalisable pour l'estimation du facteur de déformation. Il est établi que la meilleure approche pour cette tâche repose sur l'utilisation des primitives cepstrales d'ordre inférieur. Plusieurs évaluations sur des bases de données standard ont été conduites afin d'illustrer les améliorations de performances et les difficultés de perception rencontrées avec la NLCV.
La NLCV emploie un paramètre unique pour représenter les caractéristiques du locuteur. Par conséquent, cette technique ne permet pas d'atteindre les performances des autres méthodes d'adaptation reposant sur des transformations linéaires lorsque une grande quantité de données d'adaptation est disponible. De nombreuses techniques sont présentées dans cette thèse afin de combiner des techniques d'adaptation reposant sur des modèles comme par exemple la régression linéaire par maximum a posteriori structurel avec contraintes (RLMAPSC) avec la NLCV. Une telle technique propose d'utiliser la NLCV comme une transformation préliminaire sur le nœud racine de l'arborescence d'un système à RLMAPSC. Ainsi, en utilisant également une adaptation rapide, les performances s'améliorent lorsque la quantité de données d'adaptation disponibles croît. Ces techniques, bien que développées pour la SPET, peuvent aussi être employées pour la RAP. Il a ainsi été constaté que ces méthodes conduisent à des améliorations pour la RAP, surtout sous certaines conditions comme dans des environnements bruyants. Par ailleurs, d'autres améliorations pour l'adaptation rapide sont également proposées comme un terme correctif pour la NLCV, une NLCV à base de transformations multiples et utilisant des classes de régression, et une NLCV avec une distribution a priori pour l'adaptation basée sur la RLMAP non structurel. Ces techniques ont conduit à des améliorations pour à la fois la RAP et la SPET. En outre, quelques scénarios particuliers, comme des évaluations utilisant de la parole interlinguale, de la parole de personnes de sexes opposés, de la parole d'enfants ou de la parole bruitée sont présentées, pour lesquelles les méthodes d'adaptation rapide décrites dans cette thèse sont extrêmement bénéfiques. Enfin, la plupart de ces méthodes seront publiées en tant qu'extensions pour le logiciel libre HTS.

Mots-clés Normalisation de la longueur du conduit vocal, coefficients cepstraux sur l'échelle de Mel généralisée, transformations bilinéaires à base de filtres déphaseurs, adaptation rapide de primitives, synthèse de la parole statistique paramétrique reposant sur des MCC, reconnaissance automatique de la parole à base de MCC, méthode de modélisation et d'adaptation unifiée pour la RAP, espérance-maximisation, transformations de modèles, régression linéaire par maximum a posteriori structurel avec contraintes, régression linéaire de vraisemblance avec contraintes

Contents

List of Figures

List of Tables

Glossary

Technical acronyms in this thesis are exhaustively listed below.

VTLN	vocal tract length normalization
HMM	hidden Markov model
H(S)MM	hidden (semi-)Markov model
ASR	automatic speech recognition
TTS	text to speech
HTS	HMM-based text to speech synthesis system
MGCEP	mel generalized cepstrum
MCEP	mel-cepstrum
BNDAP	band aperiodicity
ML	maximum likelihood
MAP	maximum a posteriori
MLLT	maximum likelihood linear regression based transformation
MLLR	maximum likelihood linear regression
CMLLR	constrained maximum likelihood linear regression
CSMAPLR	constrained structural maximum *a posteriori* linear regression
MAPLR	maximum a posteriori linear regression
CMAPLR	constrained maximum a posteriori linear regression
CLSA	cross-lingual speaker adaptation
MOS	mean opinion score
EM	expectation-maximization
MCD	mel-cepstral distortion
MGE	minimum generation error
MSD	multi-space distribution

pdf	probability density function
STRAIGHT	speech transformation and representation using adaptive interpolation of weighted spectrum
EMIME	effective multilingual interactions in mobile environments

1 Introduction

In recent years, there has been a big boom in the smart phone market especially with interactive voice applications such as automatic dialogue systems available on these devices, personalized speech-to-speech translation is emerging as an important area of research. Success of these technologies mainly relies on the ability to personalize the voice characteristics in the underlying text-to-speech synthesis (TTS) system. The particular area of voice transformation research has much wider impact such as applications in medical (e.g. Voice generation systems for patients with throat cancer), security (e.g. voice biometrics) and entertainment (e.g. dubbing and translation of videos across languages) industries. Since these systems involve personalization of output speech, it is crucial to these kind of applications that the speaker characteristics are induced into the output speech from the very first utterance spoken by a speaker. Thus, speaker characteristics need to be estimated from very little adaptation data. Towards this end, this research work aims to perform rapid speaker adaptation for statistical parametric speech synthesis using vocal tract length normalization (VTLN), which is a widely used technique in automatic speech recognition (ASR). Another important focus of this research is to unify the techniques used in ASR and TTS, which are the integral parts of a personalized speech-to-speech translation system. The techniques developed for TTS are also investigated to be effectively used in ASR for further improvements on the state of the art.

1.1 Motivation

The advent of statistical parametric speech synthesis has considerably reduced the gap between automatic speech recognition (ASR) and text to speech synthesis (TTS). The hidden Markov model (HMM) based ASR and TTS helps to unify the two, otherwise very diverse areas of speech technology. This in turn paves the way for developing new techniques for one domain which could be easily adopted to the other. Even the techniques that show promising results in ASR, could be expected to perform well in statistical parametric synthesis.

Most of this research was performed under the "Effective multilingual interactions in mobile environments" (EMIME) personalized speech-to-speech translation project. The EMIME

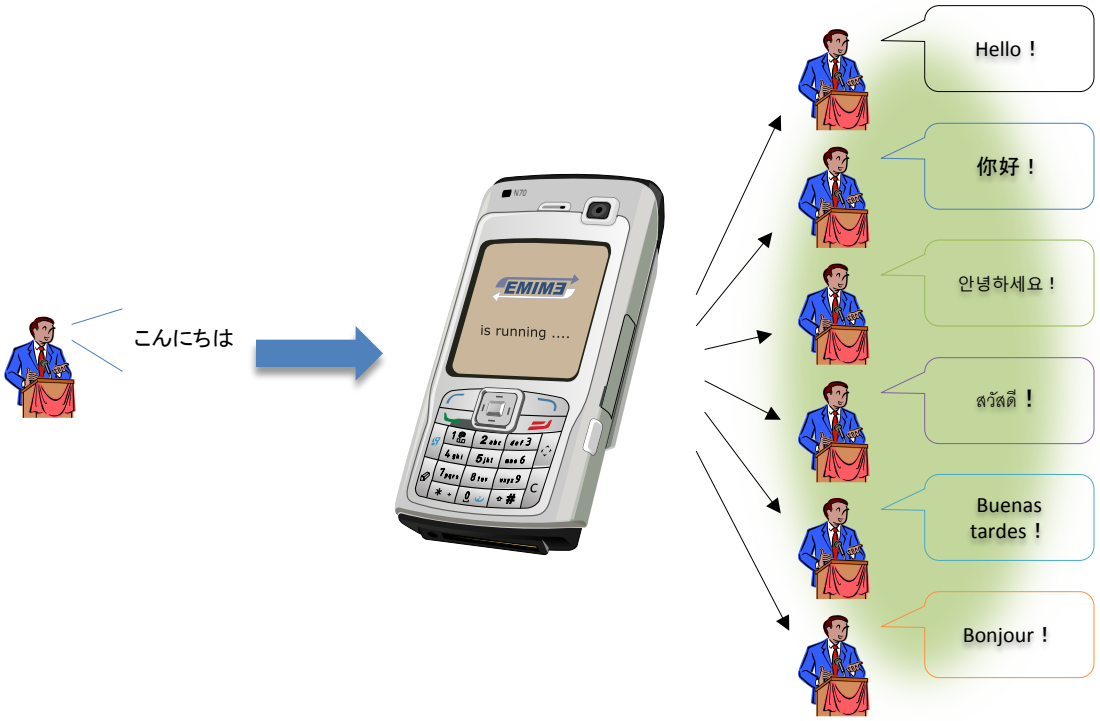


Figure 1.1 – EMIME personalized speech-to-speech translation system. This figure is modified from one of the figures of the Nagoya Institute of Technology.

project was a European Community's Seventh Framework Programme (FP7/2007-2013). The project aimed to recognize an input speech utterance from a user, translate it to a target language and then synthesize the speech in the target language with the speaker characteristics matching the input user. Hence, the output speech sounds like the same user speaking another language (shown in Figure 1.1). EMIME helps people to have conversations in languages they don't even understand. The idea is to crush the language barriers and allow people to interact more freely without even learning one another's language. It is definitely helpful to have such a system on a mobile device especially during travel to countries where a language that a visitor cannot understand is spoken. This project had many collaborating research groups from different parts of the world, addressing different aspects of the system.

The use of unified models in speech-to-speech translation represents a particularly attractive paradigm since it provides a natural mechanism for speaker-adaptive TTS by employing the same speaker dependent transforms learned from ASR, while offering further efficiency with respect to computation and memory (see for eg. [King et al., 2008]). There are numerous challenges present in developing such models. In particular, despite the common underlying statistical framework, HMM-based ASR and TTS systems are generally very different in their implementation.

The numerous challenges to consider include cross-lingual speaker adaptation, rapid speaker

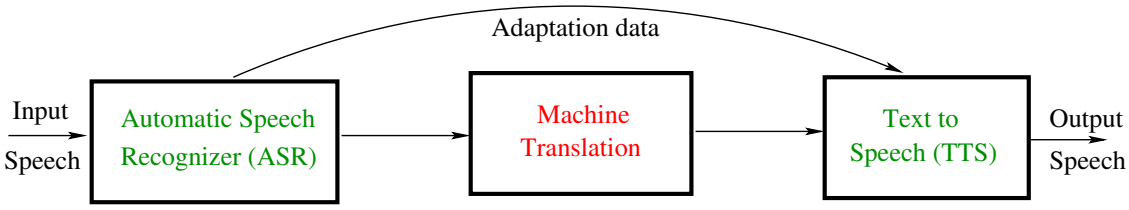


Figure 1.2 – Components of a speech-to-speech translation system.

adaptation, quality and accuracy of the output speech, and perception of the speaker identity in another language. Apart from cross-lingual aspects of the project, an equally demanding feature of the project is rapid speaker adaptation. This work looks into bringing in the target speaker characteristics even from the very first utterance of the user. This is an important feature for a personalized speech-to-speech translation system, where the user expects to hear the output speech in at least the same gender (as her/his speech) from the system.

1.2 Objective

The major themes involved in this research are:

1. Rapid Speaker adaptation.
2. Unified modeling for ASR and TTS.
3. Joint ASR and TTS features and feature adaptations.
4. Unifying the theory and practice for speaker normalization techniques.
5. Robust speaker adaptation.

The main objective of this research is to unify the features and the rapid adaptation techniques (like VTLN) used in the ASR and TTS components (shown in Figure 1.2) of the personalized speech-to-speech translation system. This unification facilitates the use of similar techniques across both systems with minor modifications. The improvements made in a domain can also be directly adopted to the other. One of the important features of a personalized speech-to-speech translation system is to bring in target speaker characteristics in the output with even a single input sentence from the user. This requires rapid speaker adaptation techniques in TTS. As mentioned earlier VTLN is a widely used feature adaptation technique in ASR, that could be implemented in statistical parametric speech synthesis. This research aims to implement VTLN for statistical parametric speech synthesis and further make improvements to the rapid adaptation scheme in combination with maximum likelihood linear transformation (MLLT) based adaptation techniques. The techniques thus developed should give improvements to the state of the art rapid adaptation in ASR as well.

1.3 Summary of Contributions

The research presented in this thesis, and first published as Saheer et al. [2010c], represents the first successful application of VTLN in statistical parametric speech synthesis to transform target voices. The contributions of this work include:

– Mel-generalized cepstrum (MGCEP) features are used with the bilinear transform based VTLN. Warping parameter estimation is carried out with ML optimization over a grid search. It is shown that VTLN brings in some speaker characteristics and provides additional improvements to constrained maximum likelihood linear regression (CMLLR), especially when there is a limited number of adaptation utterances [Saheer et al., 2010c].
– The divergence in theory and practice for VTLN implementation is analyzed in detail and different solutions are proposed to unify the two [Saheer et al., 2010b, 2012a].
– An efficient implementation of VTLN using expectation maximization with Brent's search is proposed to improve the time and space complexity of the warping factor estimation and to embed this estimation in the HMM training [Saheer et al., 2010a, 2012a].
– Different techniques are proposed to combine VTLN with more powerful model-based linear transformations like constrained structural maximum a posteriori linear regression (CSMAPLR). These techniques include using VTLN as a prior with or without structure in CSMAPLR and using VTLN as a parent transform for CSMAPLR. These techniques help to harness the power of VTLN as a rapid adaptation technique and scale it to the performance of other powerful transformations with the availability of more adaptation data. The techniques proposed in this thesis are shown to be beneficial for both TTS and ASR [Saheer et al., 2012c].
– Improvements on VTLN with the implementation of a bias term can enhance performance.
– A set of experiments also demonstrate the special situations where VTLN can be most effective, which includes the cross gender transformations, age transformations, noisy conditions, or cross-lingual transformations [Saheer et al., 2012b].

1.4 Thesis Outline

This thesis is organized as follows. The motivation, objective and contributions of this work were briefly summarized in this chapter. Chapter 2 compares the statistical parametric speech synthesis and recognition in the spirit of the unification theme of this research and also gives a general background on VTLN as a rapid adaptation technique in both ASR and TTS. Chapter 3 describes the vocal tract length normalization being formulated as an all-pass transform based feature transformation for the MGCEP features. The warping factors are being estimated using a grid search technique with the maximum likelihood criteria. The diversity in theory and practical implementation of VTLN is handled in Chapter 4. This chapter also proposes different schemes to overcome these differences and have a unified VTLN modeling. A more efficient implementation of VTLN using the expectation maximization algorithm is presented in Chapter 5. VTLN can now be represented as an equivalent model adaptation technique and Brent's search based optimization is performed on the sufficient statistics similar to the

model adaptation techniques. VTLN is combined with model transformations like CSMAPLR in order to scale the performance of VTLN when an adequate amount of adaptation data is available. The details are presented in Chapter 6 along with improvements in VTLN adaptation using the bias transformation. Finally, VTLN is evaluated in some special conditions where the constraints of VTLN as a speaker adaptation can be useful. The details of cross-gender, cross-age, cross-lingual and noisy environment evaluations are presented in Chapter 7. Chapter 8 discusses the conclusions of this work along with a summary of major contributions and future directions.

2 Background

The work presented in this thesis uses an HMM-based statistical parametric speech synthesis and performs evaluations on its ASR counterpart as well with the aim of unification of adaptation techniques for ASR and TTS. This chapter describes the general background needed for this work.

2.1 HMM based Recognition vs Synthesis

Recent advances in the field of statistical parametric speech synthesis by Black et al. [2007], have considerably reduced the gap between basic techniques used in automatic speech recognition (ASR) and text to speech (TTS). Feature types, feature dimensionality, duration and pitch modeling are a few of the key differences between the recognition and synthesis models as shown by Dines et al. [2010]. To augment the ASR models, speech synthesis also uses a duration model by way of the hidden semi Markov models (HSMM). Even with these differences in place many of the techniques used to improve speech recognition quality can be successfully applied to speech synthesis as well [Yamagishi et al., 2009a]. The main interest here is to unify the features used for ASR and for TTS [Dines et al., 2009a]. In particular, the interest is in the use of ASR based adaptation to control the characteristics of a synthesized voice.

HMM-based Speech Synthesis System (HTS) by Zen et al. [2004] models spectrum, F0 and duration simultaneously in the unified framework of HSMM. The details of the system are shown in Figure 2.1. In the training stage, the output vector of the HSMM consists of a spectrum part, a band aperiodicity (bndap) part and an F0 part. A multi-space probability (MSD) HSMM is used to model one dimensional continuous and discrete valued F0 patterns without heuristic assumption. In the synthesis stage, an arbitrarily given text is converted to a context-dependent label sequence. A sentence HSMM is constructed by concatenating corresponding HSMM models. A state sequence that maximizes the probability for the given sentence is determined. Then a speech parameter vector sequence is generated for this state sequence by using a maximum likelihood inference. Finally, the speech waveform is generated from the speech

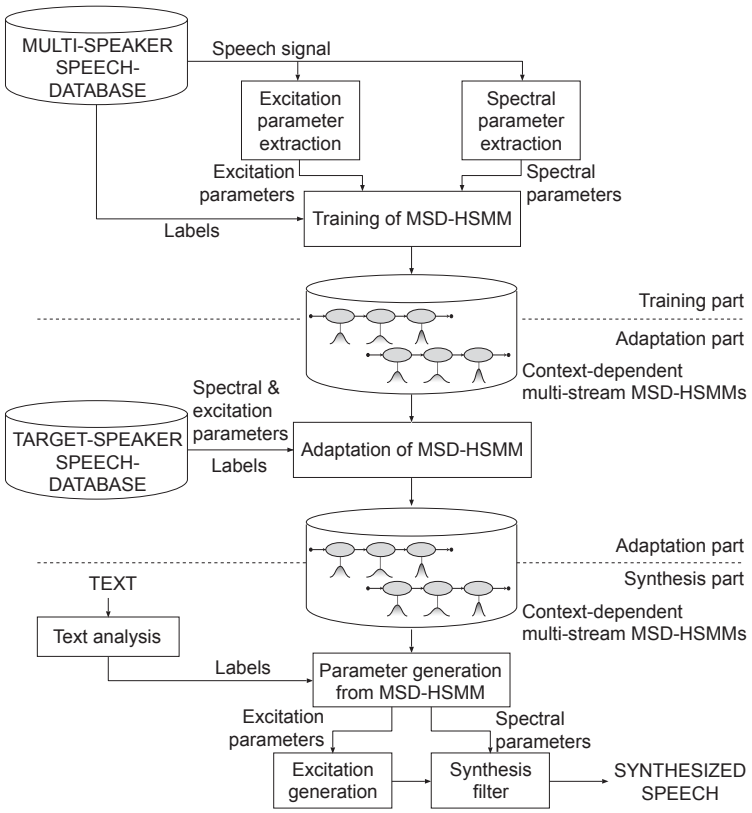


Figure 2.1 – Overview of HSMM based statistical parametric speech synthesis. Figure courtesy Dines et al. [2010]

parameter vector sequence. Adaptation techniques are used in the same way in TTS and ASR. Speaker independent models are built using the adaptation techniques that remove speaker characteristics. During synthesis, models are adapted for synthesizing speech of a particular speaker using the corresponding adaptation data.

The steps involved in the recognition stage of an ASR system is shown in the Figure 2.2. There is a possibility that the adaptation data from the target speaker does not have transcriptions. In such a case (unsupervised speaker adaptation), a two pass recognition technique is used, where, the transcriptions in the first pass are generated using a trained model. These transcripts are used along with the feature vectors to adapt this model to the target speaker and perform recognition using the adapted model in the second stage. Following the discussions by Dines et al. [2010], the main differences in the implementations of a typical HMM-based ASR and TTS are summarized in the Table 2.1s.

There are also differences in the basic features used in the two techniques as discussed in the next section.

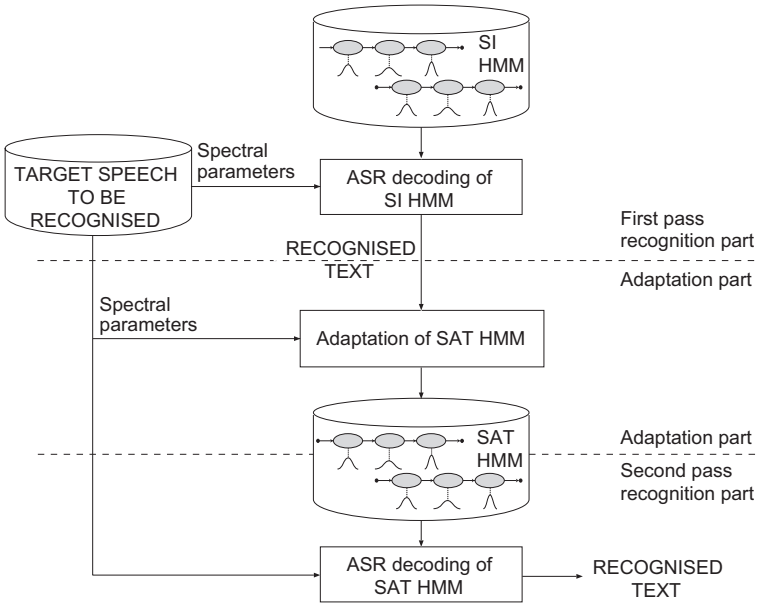


Figure 2.2 – Overview of HMM based recognition system. Figure courtesy Dines et al. [2010]

2.1.1 MGCEP and MFCC features

Mel-Generalized Cepstral Coefficients (MGCEP) [Tokuda et al., 1994b] are one of the widely used features for statistical speech synthesis. As explained by Tokuda et al. [1994b], Linear prediction is a generally accepted method for obtaining all-pole representation of speech. However, in some cases, spectral zeros are important and a more general modeling procedure is required. Although cepstral modeling can represent poles and zeros with equal weights, the cepstral method with a small number of cepstral coefficients overestimates the bandwidths of the formants. This can be overcome by a generalized cepstral analysis technique. The generalized cepstral analysis method is viewed as a unified approach to the cepstral method and the linear prediction method, in which the model spectrum varies continuously from all-pole to cepstral. There are two parameters (α and γ) that determine the behavior of the feature extraction and could be optimized for a particular task. The generalized cepstral coefficients are equivalent to the cepstral and autoregressive (AR) coefficients when the analysis (compression) parameter (γ), equals 0 and -1 respectively. The warping parameter(α) determines the frequency warping of the cepstra and is based on bilinear transform of an all-pass filter. The steps involved in the MGCEP feature analysis are shown in Figure 2.3. Note that MGCEP is different from the traditional MFCC features most commonly used in ASR, in particular, there is no filterbank analysis. A generalized log function is applied on the squared magnitude spectrum and the mel-warping is performed in the cepstral domain using the all-pass filter based bilinear transform.

The generalized logarithmic function is a natural generalization of the logarithmic function

Table 2.1 – Differences between ASR and TTS shown by Dines et al. [2010]

Attributes	ASR	TTS
General		
Lexicon	CMU	UniSyn
Phone set	CMU (39 phones)	GAM (56 phones)
Acoustic parametrization		
Spectral analysis	fixed size window	STRAIGHT (F0 adaptive window)
Feature extraction	filter-bank cepstrum ($\Delta + \Delta^2$)	mel-generalized cepstrum ($+ \Delta + \Delta^2$) + log F0 + bndap ($+ \Delta + \Delta^2$)
Feature dimensionality	39	120 + 3 + 15
Frame shift	10ms	5ms
Acoustic modeling		
Number of states per model	3	5
Number of streams	1	5
Duration modeling	transition matrix	explicit duration distribution (HSMM)
Parameter tying	phonetic decision tree (HTK)	shared decision tree (MDL)
State emission distribution	16 component GMM	single Gaussian pdf
Context	triphone	full (quinphone + prosody)
Training	2-pass system (ML-SI & ML-SAT)	Average voice (ML-SAT)
Speaker adaptation	CMLLR	CMLLR or CSMAPLR

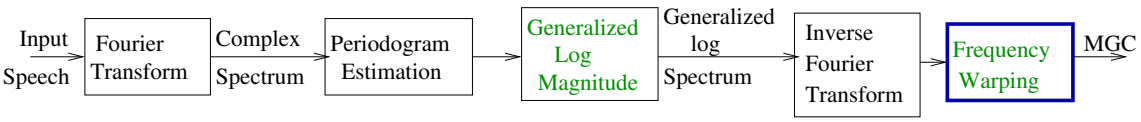


Figure 2.3 – Key stages of mel-generalized cepstral analysis. UELS and cepstral truncation are omitted. Spectral warping is implemented as a linear transformation in the cepstral domain.

with a parameter γ as shown in Figure 2.4:

$$s_\gamma(\omega) = \begin{cases} \frac{\omega^\gamma - 1}{\gamma}, & 0 < |\gamma| \leq 1, \\ \log\omega, & \gamma = 0. \end{cases} \tag{2.1}$$

where, s is the generalized function and ω represents the magnitude spectrum. More details on the MGCEP features especially the γ parameter formulation is given in Appendix A. The warping function used in this feature extraction technique is the bilinear transform based all-pass warping with warping parameter α. This frequency transformed generalized cepstrum has frequency resolution similar to that of the human ear with an appropriate choice of the value of the warping parameter(α). Hence, it is expected that the mel-generalized cepstral coefficients are useful for speech spectrum representation. The different systems represented by various combinations of the values of α and γ are shown in Figure 2.5.

Mel-frequency cepstral coefficients (MFCCs) are one of the most commonly used features in ASR. These are different from the MGCEP features because of the presence of the filter bank

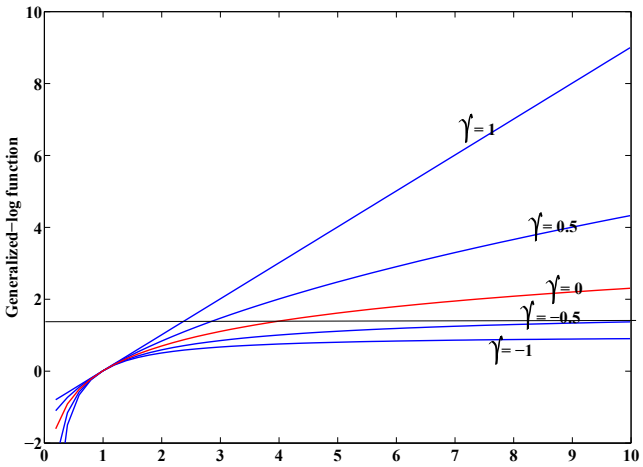


Figure 2.4 – Generalized Log Function

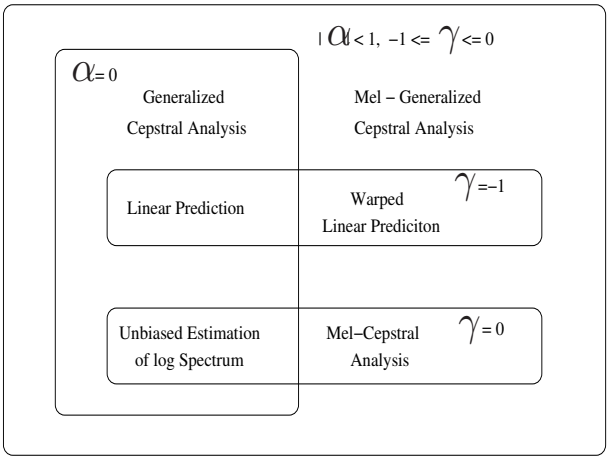


Figure 2.5 – MGCEP Analysis. Figure based on Tokuda et al. [1994b]

smoothing. The steps involved in this feature extraction technique are shown in Figure 2.6. The magnitude spectrum of the speech is passed through a set of overlapping triangular bins uniformly spaced in the mel-scale. A log of the smoothed spectrum followed by DCT function results in the cepstral coefficients. The MFCC tends to remove the speaker specific or excitation parameter information in the cepstra in order to maximize the information regarding the speech sound represented by the features which is what is required in a recognition system. The main differences between MFCC and MGCEP are illustrated in Table 2.2.

The spectral model based on the mel-generalized cepstral representation also includes the criterion used in the unbiased estimation of log spectrum (UELS). It is shown [by Tokuda et al., 1994b] that the minimization of the criterion is equivalent to the minimization of the mean square of the linear prediction error. When the γ parameter is zero, the proposed method is similar to the cepstral representation and UELS can be directly used to minimize the spectral

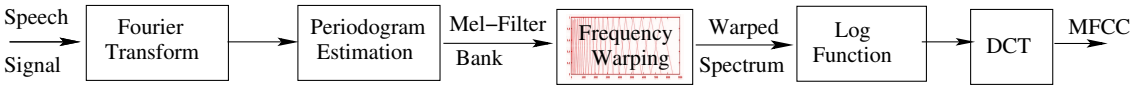


Figure 2.6 – Key stages of MFCC Analysis.

Table 2.2 – MFCC vs MGCEP

MFCC	MGCEP
filter bank smoothing	no binning
spectral (frequency) warping	cepstral warping
log cepstra	generalized log cepstra
13 dimensional static cepstra	39 dimensional static cepstra
not easy to reconstruct spectrum	revertible spectral representation

distortion. When $\gamma = -1$, the method represents a linear prediction approach and the UELS minimization criterion becomes a set of linear equations in the linear prediction method. With other values of γ, different spectral analysis techniques can be represented and minimization becomes non-linear. Although the method involves a non-linear minimization problem, it can easily be solved by an iterative algorithm [as shown by Tokuda et al., 1998]. The convergence is quadratic and typically a few iterations are sufficient to obtain the solution. The stability of the obtained model solution is also guaranteed. As a result, this method can be viewed as a unified approach to speech spectral analysis, which includes several speech analysis methods.

The block diagram representing the steps involved in MGCEP, omits certain details such as UELS and cepstral truncation. UELS could be avoided in the mel-cepstral (MCEP) features where the γ parameter is set to zero and a log spectrum replaces the generalized log spectrum. Since optimization of γ is an optimization problem in itself, most of this work uses MCEP features ($\gamma = 0$). A few experiments have also been performed with MGCEP features using $\gamma = 3$, which represents the cube root of the spectrum similar to compression in perceptual linear prediction (PLP) features.

Despite these differences in place, both ASR and TTS can still use a unified model with similar adaptation techniques.

2.2 VTLN as a rapid feature adaptation technique

Speaker adaptation is a technique for transforming the model parameters to match the speaker characteristics of a target speaker. Speaker adaptive training helps to build improved speaker independent models by transforming the model parameters and removing speaker characteristics for each speaker in the training data. The most common adaptation techniques are MLLR (Maximum Likelihood Linear Regression), CMLLR (Constrained MLLR), SMAPLR (Structural Maximum A Posteriori Linear Regression) and CSMAPLR (Constrained SMAPLR). Speaker normalization, on the other hand, transforms the feature vectors rather than the

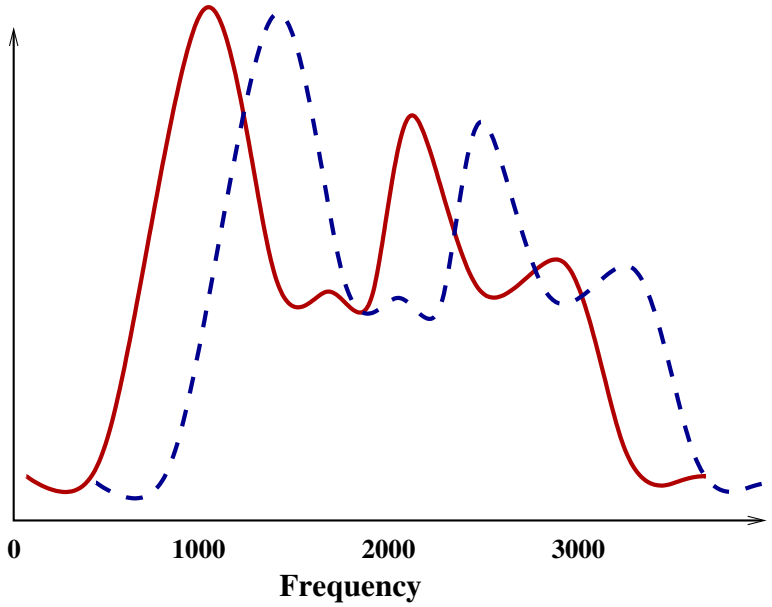
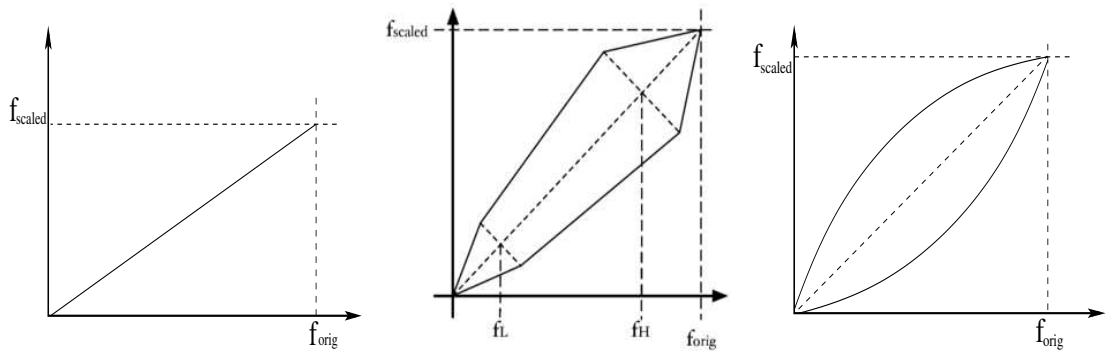


Figure 2.7 – Spectral transformation in VTLN

model parameters. Feature transformation can be shown to be analogous to model transformation as shown by Pitz and Ney [2005]. Usually, speaker adaptation techniques perform affine transformations on the mean and variance of the probability density functions of the HMM states. This can be accomplished to some extent with normalization techniques like vocal tract length normalization (VTLN). The main advantage of speaker normalization is that the number of parameters to be estimated from the adaptation data is generally smaller compared with the standard model based adaptation techniques. Hence, adaptation can be carried out with very little adaptation data.

VTLN is inspired from the fact that the vocal tract length varies across different speakers. This length varies from around 18 cm in males to around 13 cm in females. Though not significantly, the vocal tract length even varies within a speaker when producing different sounds. The formant frequency positions are inversely proportional to the vocal tract length. This causes an approximate variation of 25% in the formant center frequencies among speakers. Hence, the feature vectors extracted from the speech of different speakers can be normalized to represent an average vocal tract. A sample transformation of the spectra using VTLN is illustrated in Figure 2.7.

The components involved in this technique are:
– Warping function (linear, piecewise linear, non-linear, bilinear, etc.)
– Warping factor (α for bilinear transform)
– Optimization criteria (MAP, ML, MGE, etc.)
The different types of warping functions usually used in VTLN is shown in Figure 2.8. Optimal parameters for the warping function referred to as warping factors are estimated based on a



a) Linear b) Piece-wise Linear b) Bilinear

Figure 2.8 – Warping functions for VTLN

pre-determined optimization criterion. The warping factors are usually selected from a grid of available values. This technique involves high time and space complexity due to the extraction of features for each speaker using each warping factor in the grid.

VTLN is a widely used technique in ASR, made particularly attractive due to its simplicity and robustness. While such advantages may equally be of interest in TTS, there have been very few attempts to date towards this end. This section presents some details of the earlier work in this field, and hence motivates the research presented in this work.

2.2.1 VTLN for Automatic Speech Recognition

Spectral transformations are closely tied to the underlying feature analysis technique, hence, implementation of VTLN cannot be easily separated from the feature extraction stage. The use of VTLN could be dated back to Wakita [1977] where the vocal tract length estimated using formant positions (referred as the resonance frequencies of the tube model) during feature extraction was used to normalize the formant frequencies. These normalized frequencies have eliminated inter-speaker variations and were used to recognize vowels. It was shown by Acero and Stern that the frequency normalization using all-pass transforms accomplishes a non-linear frequency warping of the cepstra using a matrix multiplication as shown by Oppenheim [1972]. The warping factor was selected through minimization of VQ error and used to warp the linear prediction cepstral coefficient (LPCC) which resulted in an average 10% reduction for the recognition error rate. The detailed study of this work on how the frequency normalization can compensate for a new user and acoustic environment was presented by Acero [1993]. A cascade of bilinear transforms representing mel-scale warping and frequency normalization was used on the LPCC to close the gap between speaker dependent and speaker independent models. It was also suggested to simplify the matrix representation of bilinear transform by ignoring the terms with power of α (the warping factor) greater than one.

A parametric method of normalization which counteracts the effect of varied vocal tract length in recognition systems was proposed by Eide and Gish [1996]. It was shown that non-linear

scaling performed better than linear scaling and warping improves recognition performance especially when small amounts of data was available from the speakers.

An initial implementation of VTLN for ASR by Lee and Rose [1998] (initially presented as Lee and Rose [1996]) was carried out in the framework of MFCC features. The MFCC features are usually estimated by smoothing the log magnitude spectrum using a bank of filters with equal spacing in the mel-frequency scale. An efficient vocal tract length warping was implemented by modifying the filterbank computation. The optimal warping factor for a piecewise linear warping function was estimated using a maximum likelihood (ML) based optimization criterion given by:

$$\hat{\alpha}_s = \underset{\alpha_s}{\operatorname{argmax}}\, p(\boldsymbol{x}_{\alpha_s} \mid \boldsymbol{\Theta}, w_s) \tag{2.2}$$

where $\boldsymbol{x}_{\alpha_s}$ represents the warped feature vectors using the warping factor α_s. s represents the target speaker and α_s is the warping factor for this speaker. $\boldsymbol{\Theta}$ represents the model and w_s represents the transcription corresponding to the data from which the features are extracted for speaker s. $\hat{\alpha}_s$ represents the best warping factor for the same speaker. This expression for feature transformation ignores the Jacobian normalization term usually implemented as the determinant of the transformation matrix. Simultaneously updating the features and calculating the likelihood scores is not a very consistent method for performing VTLN. The re-estimation of spectrum from the cepstral parameters to extract each set of warped feature vectors is not an ideal implementation of VTLN.

It was shown by Pye and Woodland [1997] that VTLN can be used for speaker adaptive training (SAT) to improve the speaker independent models along with maximum likelihood linear regression (MLLR) transforms. A filter bank based normalization approach was used that gave additive improvements to the MLLR transforms. Another interesting approach was 3-D Viterbi decoding using the 3-D trellis space composed of input frames, HMM states and warping factors by Fukada and Sagisaka [1998]. Using these time varying frequency warping factors improved the spontaneous recognition performance. Different implementations of VTLN warping was compared by Uebel and Woodland [1999] and it was shown that VTLN could have additive improvements with MLLR, but, not so much with constrained MLLR (CMLLR). There were a lot of different efforts within the research community across the world to improve the VTLN techniques and it is almost impossible to mention every effort here. The notable examples include improving the VTLN training and warping factor estimation by Welling et al. [1999] or the implementation of VTLN as a band diagonal transformation by Afify and Siohan [2000] who termed this as a constrained form of MLLR [1]. The fast VTLN approach by Welling et al. [1999] requires multiple acoustic models (normalised, unnormalised and gaussian models for scale factor estimation), estimation of multiple feature vectors with different scaling and still does not perform as well as the data with labels (multi-pass strategy). In this work, the warping factor estimation requires a single sentence which could be just a

1. Although the name sounds similar to the CMLLR speaker adaptation technique, here the reference is to bilinear transform based VTLN expressed as a model transformation.

phrase or even a word or sub-word unit and performs better than the fast VTLN approach by Welling et al. [1999].

The all-pass transform can be used to approximate most commonly used transformations in VTLN [McDonough, 2000, Uebel and Woodland, 1999]. The all-pass transform based VTLN introduced a broader range of possibilities for transformation of the spectra in the ASR domain. The bilinear transform of a pole-zero pair, which is the simplest form of the all-pass transforms (shown in Figure 3.1), has only a single variable α as the warping factor. This parameter is representative of the ratio of the vocal tract length of the speaker to an average vocal tract length. The terms warping factor and 'α' refer to the same parameter and are used interchangeably throughout this work.

A conformal mapping based cepstral warping using analytic functions for speaker adaptation was presented by McDonough et al. [1996]. It was presented by McDonough [2000] that the all-pass transforms based mapping for frequency normalization can be generalized to rational all-pass transforms (RAPT) and sine-log all-pass transforms (SLAPT). These sophisticated mapping techniques enable separate transformation for each formant of the speech spectrum which can be seen as a perturbation atop the frequency transformation induced by bilinear transforms.

It was proven analytically by Pitz [2005] that vocal tract length normalization (VTLN) equals a linear transformation in the cepstral space for arbitrary invertible warping functions. The transformation matrix for VTLN was explicitly calculated for three commonly used warping functions (piece-wise linear, quadratic and bilinear warping). Based on some general characteristics of typical VTLN warping functions, a common structure of the transformation matrix was derived that is almost independent of the specific functional form of the warping function. By expressing VTLN as a linear transformation it was possible to take the Jacobian determinant of the transformation into account for any warping function. The transformation matrices were found to be diagonally dominant and it was postulated that they could be approximated using quindiagonal matrices. The effect of considering the Jacobian determinant on the warping factor estimation was also studied.

Expressing warping as a linear VTLN transformation of the cepstra and using auxiliary functions similar to the ones used in CMLLR were proposed by Kim et al. [2004]. It was shown that linear VTLN performance was comparable to the conventional VTLN approach of coding the data with multiple warp factors. VTLN was expressed as a linear transformation of the MFCC features by Umesh et al. [2005] after considering the idea of quefrency limitedness of filterbank smoothing and performing the linear transformation in the discrete cepstral domain. The mel-scale warping is also separated from the filter bank and implemented along with the VTLN warping as linear cepstral transformation. This enables the use of any warping function and results in the linear cepstral transformations for MFCC features with comparable performances to conventional VTLN warping. While this was sinc-interpolation of the spectrum, a cosine interpolation based linear VTLN transformation was presented by Panchapagesan and Alwan

[2009] which performed the warping using IDCT matrices. Based on the work of Umesh et al. [2005], Sanand and Umesh [2012] proposed a linear transformation based VTLN that replicates the conventional VTLN approach without modifying the signal processing steps of MFCC features. The importance of (accuracy of) the feature extraction in using VTLN was further emphasized by [Garau, 2008] while using pitch-synchronous acoustic features extracted using STRAIGHT tool.

2.2.2 VTLN in Text to Speech Synthesis & Voice Conversion

Though widely used in ASR, VTLN is a comparatively less exploited technology in TTS. Earlier attempts at using VTLN for concatenative speech synthesis by Sündermann [2008] had the advantage of performing voice conversion without needing to use corresponding time frames from the source and target speakers. This method was particularly useful in cross-language voice conversions where natural time alignments across speech from different languages could not be obtained. Sündermann [2008] performed voice conversion by clustering the phoneme classes of source and target speakers separately, and then statistically mapping the phoneme classes across speakers.

Voice conversion is the process of automatic transformation of a source speaker's voice to that of a target speaker. State-of-the-art voice conversion algorithms employ two common stages: Training and transformation. The voice conversion system gathers information from the source and target speaker voices and automatically formulates voice conversion rules at the training stage. The transformation stage employs the conversion rules to modify the source voice in order to match the characteristics of the target voice. Since VTLN is a simple language independent technique to alter the speaker characteristics, it can be easily developed as a voice conversion tool.

Voice conversion techniques are mainly useful in concatenative synthesis (as shown by Sündermann [2008]), where the transformation of the speaker characteristics is a difficult task due to the absence of a model or proper speech parametrization. This method of voice conversion is non trivial and involves tuning of the distance measure and the minimization criterion, and also requires parameter smoothing techniques. With the advent of HMM-based statistical speech synthesis, voice conversion techniques, including VTLN, can be implemented in a more subtle manner [Yamagishi et al., 2009a].

2.2.3 VTLN for HMM-based Synthesis

There have been earlier attempts to use VTLN for building better speaker adaptive models [Hirohata et al., 2003] for statistical speech synthesis. There is no published research work on using VTLN for statistical parametric speech synthesis.

Following the work in this thesis (initially presented in [Saheer et al., 2010c]), a formant based VTLN transformation was attempted by Zhuang et al. [2010]. A piecewise linear warping

function was estimated by mapping the first four formants of the long vowels of a source and target speaker. There were a number of drawbacks to this technique such as including accuracy of the formant estimation, using only a few vowels for mapping, and warping function being specific to a source-target pair, etc. Still, the method showed additional improvements to the MLLR adaptation technique, especially in cross-gender scenario. The research presented in this dissertation gives a more refined implementation of VTLN for statistical parametric speech synthesis and proposes techniques to improve the performance by combining VTLN with other model transformation techniques.

2.3 Other Linear transformations

There are other linear transformations available for speaker adaptation, viz., Maximum likelihood linear regression (MLLR), Constrained MLLR (CMLLR) [Gales, 1998], and Constrained structural maximum a posteriori linear regression (CSMAPLR) [Yamagishi et al., 2009a]. VTLN performs speaker adaptation similar to what these model transformations achieve when there is very little adaptation data. But, as such due to the limitation of number of available parameters in VTLN, it is not possible to scale to the performance of other linear transformations with the availability of more adaptation data. This will be dealt with in detail in the later chapters. Even for the initial implementation as a feature transformation using grid search, it is possible to combine VTLN with other linear transformations like CMLLR to further improve the performance.

MLLR transforms proposed by Leggetter [1995], estimate a set of linear transformations for the mean parameters of a mixture Gaussian HMM system to maximize the likelihood of the adaptation data.

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \text{ and } \hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} \tag{2.3}$$

where $\boldsymbol{A}$ is the transformation matrix, $\boldsymbol{b}$ is the bias vector and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ correspond to the mean and variance of the GMM. In order to closely model data in either speaker adaptation or acoustic environment compensation the Gaussian variances should also be modified. The model variances should be reduced towards a speaker dependent system removing the inter-speaker variability. Separate unrelated transforms could be used for the means and variances [Gales, 1998] referred to as unconstrained MLLR.

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \text{ and } \hat{\boldsymbol{\Sigma}} = \boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^T \tag{2.4}$$

where, $\boldsymbol{H}$ is the transformation matrix for the variance.

Constrained MLLR (CMLLR) was first introduced by Digalakis et al. [1995]. The same transformation is applied to means and variances of the Gaussian models.

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \text{ and } \hat{\boldsymbol{\Sigma}} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T \tag{2.5}$$

This can be shown to be equivalent to the transformation of the feature vector, state observation variable, $\boldsymbol{o}(\tau)$:

$$\hat{\boldsymbol{o}}(\tau) = \boldsymbol{A}^{-1}\boldsymbol{o}(\tau) + \boldsymbol{A}^{-1}\boldsymbol{b} \tag{2.6}$$

It was shown by Gales [1998] that there exists an iterative solution to optimize this transformation matrix. CMLLR is also referred to as feature space MLLR (fMLLR) and there are even techniques proposed to improve the time complexity of the fMLLR transform estimation [Varadarajan et al., 2008] termed as Quick fMLLR. Only linear transformations based on the maximum likelihood criterion are considered in this thesis since the same criterion is used in this VTLN implementation. There could be transformations based on other criteria like the minimum generation error (MGE), the minimum phone error (MPE) or the maximum mutual information (MMI).

3 Vocal Tract Length Normalization

The conventional technique for implementation of VTLN is traversing a grid of possible warping factors and searching for the best suitable value from this grid based on a given criterion for the adaptation data. This technique is referred to as "grid search" in this work. The most common feature representation used in statistical parametric speech synthesis is the mel-generalized cepstral (MGCEP) features. Although, VTLN has been successfully used in ASR, this work is the first successful attempt to implement VTLN for statistical parametric speech synthesis. This chapter explains the details of implementing VTLN using the grid search technique on MGCEP features for HMM-based speech synthesis.

3.1 Related Work

A bilinear transform of an all-pass filter is used as the warping function with the maximum likelihood criterion on the MGCEP features in this work. Most of the background needed for this work was presented in Chapter 1. This section gives some background on the all-pass transforms. These are the same transforms used in the MGCEP features (also presented in chapter 1).

3.1.1 All-pass transforms

The bilinear transform converts any function in the s-plane to a discrete function in the z-plane. Padé's approximation helps to find the correspondence between the Laplace and the z-transforms.

$$z = e^{sT} = \frac{e^{\frac{sT}{2}}}{e^{\frac{-sT}{2}}} \approx \frac{1+\frac{sT}{2}}{1-\frac{sT}{2}} \quad \text{which implies,} \quad s \approx \left(\frac{2}{T}\right)\frac{1-z^{-1}}{1+z^{-1}} \tag{3.1}$$

Consider an all-pass filter with a pole and a zero, $H(s) = \left(\frac{a-s}{a+s}\right)$, a bilinear transform can convert

this to:

$$\frac{a-\left(\frac{2}{T}\right)\frac{1-z^{-1}}{1+z^{-1}}}{a+\left(\frac{2}{T}\right)\frac{1-z^{-1}}{1+z^{-1}}}=\frac{T(1+z^{-1})a-2(1-z^{-1})}{T(1+z^{-1})a+2(1-z^{-1})}=\frac{z^{-1}-\alpha}{1-\alpha z^{-1}} \tag{3.2}$$

with $\alpha=\frac{2-aT}{2+aT}$. Hence, the bilinear transform of a simple first order all-pass filter with unit gain can be represented as:

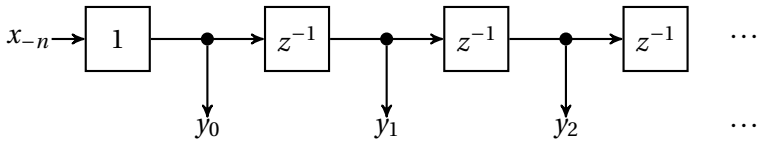
$$\psi_\alpha(z)=\frac{z^{-1}-\alpha}{1-\alpha z^{-1}}, |\alpha|<1 \tag{3.3}$$

where α is the warping factor.

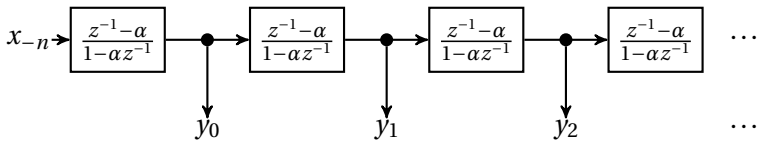
It was shown by Oppenheim [1972], a transformation that converts one discrete signal sequence into another discrete sequence preserving convolution should have the property:

$$\phi_k(z)=[\phi_1(z)]^k \text{ where, } y_k=\sum_{n=-\infty}^{+\infty}x_n\phi_{n,k} \tag{3.4}$$

ϕ_k transforms the sequence x_n to y_k. If it is a unit delay transform as shown by the lattice below [1] then $\phi_n(z)=z^{-n}$.



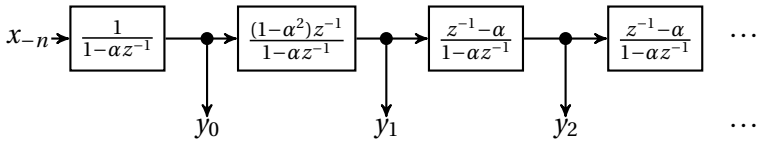
The unit delay can be replaced by a sequence of all-pass filters, $\phi_n(z)=\frac{z^{-1}-\alpha}{1-\alpha z^{-1}}$.



In order to compensate for causality and border conditions, it was shown by Oppenheim [1972] that all-pass bilinear transform for warping the spectra was in fact more complicated transformation represented by

$$\phi_k(z)=\frac{(1-\alpha^2)z^{-1}}{(1-\alpha z^{-1})^2}\left[\frac{z^{-1}-\alpha}{1-\alpha z^{-1}}\right]^{k-1} \text{ and, } \phi_0(z)=\frac{1}{1-\alpha z^{-1}} \tag{3.5}$$

1. Figures courtesy Garner [2010]



Since the transformation is based on all-pass filters, it is hereby referred to as the all-pass transformations in the rest of this thesis. This transform preserves the magnitude but shifts the phase, hence performing the frequency warping as proven below (derivation shown by McDonough [2000]). Consider the all-pass transform as a frequency transform represented as:

$$\psi_\alpha(z) = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}} = \frac{1-\alpha z}{z-\alpha} = Me^{-j\beta_\alpha(\omega)}. \tag{3.6}$$

where M is the magnitude of the transformation, which can be shown to be unity. For simplicity it can be represented as:

$$M'e^{j\beta_\alpha(\omega)} = \frac{z-\alpha}{1-\alpha z} \tag{3.7}$$

where $M' = \frac{1}{M}$, evaluating the values on unit circle, $z = e^{j\omega}$,

$$M'e^{j\beta_\alpha(\omega)} = \frac{e^{j\omega}-\alpha}{1-\alpha e^{j\omega}} = e^{j\omega}\frac{1-e^{-j\omega}}{1-e^{j\omega}} \tag{3.8}$$

The magnitude of $e^{j\omega}$ is unity and the numerator of the rational term is the complex conjugate of the denominator which implies, the quotient has unit magnitude. Thus both M and M' has unit magnitude. Hence,

$$e^{j\beta_\alpha(\omega)} = \frac{z-\alpha}{1-\alpha z} \tag{3.9}$$

In order to derive the frequency transformation, multiply the numerator $N(z)$ and denominator $D(z)$ by the conjugate of the denominator which yields:

$$\frac{N(z)}{D(z)} = \frac{z-\alpha}{1-\alpha z} \times \frac{1-\alpha\hat{z}}{1-\alpha\hat{z}} \tag{3.10}$$

Since $D(z)$ is real regardless of the value of α or z, $\arg(\psi_\alpha(z)) = \arg(N(z))$.

$$N(z) = z - \alpha|z| - \alpha + \alpha^2\hat{z} \tag{3.11}$$

Again, since z is limited to the unit circle, $|z| = 1$ and substituting $z = e^{j\omega}$ gives,

$$N(e^{j\omega}) = e^{j\omega} - 2\alpha + \alpha^2 e^{-j\omega} \tag{3.12}$$

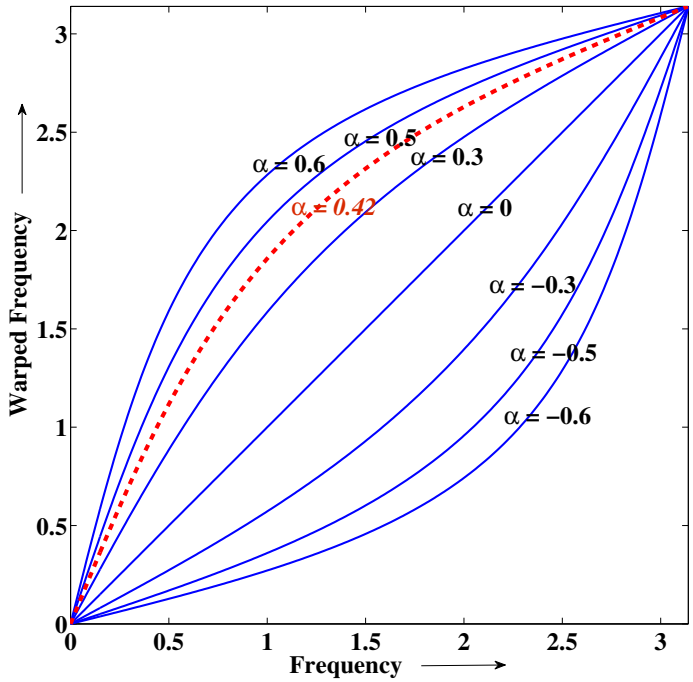


Figure 3.1 – The All-pass Transform

Using the identity $e^{j\omega} = \cos\omega + j\sin\omega$,

$$N(e^{j\omega}) = \cos\omega + j\sin\omega - 2\alpha + \alpha^2(\cos\omega - j\sin\omega) = (1+\alpha^2)\cos\omega - 2\alpha + j(1-\alpha^2)\sin\omega \quad (3.13)$$

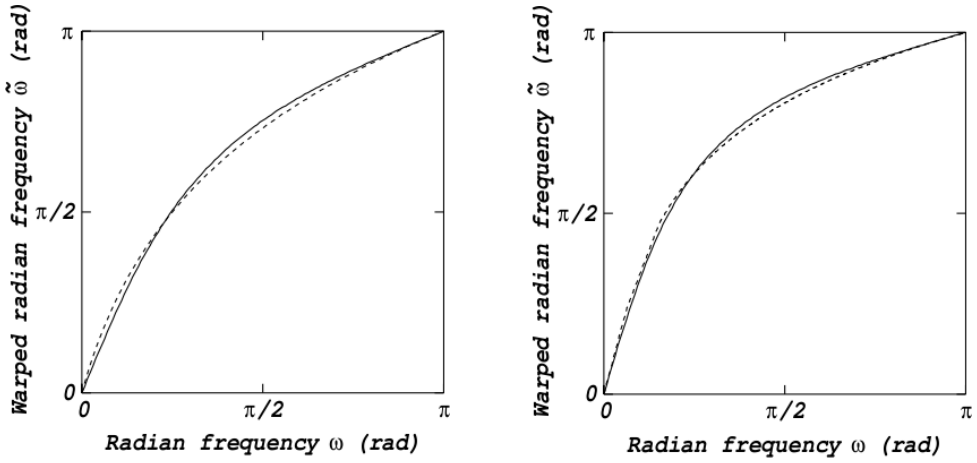
Defining the transformed angular frequency, $\tilde{\omega} = \arg(\psi_\alpha(e^{j\omega})) = \arg(N(e^{j\omega}))$, the frequency transformation achieved is defined by:'

$$\tilde{\omega} = \beta_\alpha(\omega) = \tan^{-1}\frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha}. \quad (3.14)$$

which is the phase response of an all-pass filter. This frequency warping is shown in Figure 3.1. It can be observed from the figure 3.2 that, for a specific value of $\alpha = 0.42$ or $\alpha = 0.55$, this transform can even approximate the mel-scale or bark-scale warping respectively.

Though the transformation presented in this work is the bilinear transformation of an all-pass filter, it was mostly referred to simply as a bilinear transformation in the literature. In this thesis, it is mostly referred to as an all-pass transformation. But, it should be noted that all these terms can be used interchangeably and refers to the same function. The main advantages of using all-pass transforms are:

1. It is invertible and can easily transform any discrete sequence.
2. The bilinear transform of the all-pass filter approximates to a reasonable degree the frequency domain transformations most often used in VTLN [Pye and Woodland, 1997]. It can even approximate the mel-scale frequency warping as shown by Tokuda et al.



(a) mel scale ($\alpha = 0.42$)

(b) Bark scale ($\alpha = 0.55$)

Figure 3.2 – $\alpha = 0.42$ approximates mel-frequency scale and $\alpha = 0.55$ approximates bark scale at 16kHz sampling rate. (based on Tokuda et al. [1994a])

[1994b].

3. It can be represented as a linear transformation in the cepstral domain. Hence, it can be easily combined with other transformations like CMLLR.

3.2 Grid Search Approach

The MGCEP feature extraction technique involves optimization of two parameters (namely, α and γ). The warping parameter, α, determines the frequency warping of the cepstra and is based on all-pass bilinear transforms. The frequency transformation used in MGCEP extraction is the same as the transform often used in VTLN. Hence, in this work, these two transforms are combined, and VTLN is applied at the feature extraction step.

The all-pass transformation used in MGCEP features can be represented as a linear transformation of the cepstral features. In the case of the grid search technique, this linear transformation matrix can be used to calculate the grid of frequency warped cepstral features and also for calculating the Jacobian normalization [2] factor estimated as determinant of the transformation matrix, used in maximum likelihood optimization of warping factors. The following derivation proves the linear transformation representation of the all-pass transformations in MGCEP similar to the derivation shown by [Pitz and Ney, 2005]. This linear representation will simplify the grid search warping factor estimation and further paves the way for refined implementations of VTLN using expectation maximization (EM). The m-th mel-cepstral coefficient, that is, frequency warped cepstrum, $\tilde{c}_m$ in MGCEP can be defined as

2. details given in chapter 4

$$\tilde{c}_m = \frac{1}{2\pi j} \oint_C \log X(\tilde{z})\, \tilde{z}^{m-1} d\tilde{z} \tag{3.15}$$

$$\log X(\tilde{z}) = \sum_{m=-\infty}^{\infty} \tilde{c}_m \tilde{z}^{-m} \tag{3.16}$$

Since the frequency warping is $X(\tilde{z}) = X(z)$, there is a linear transformation in the cepstral domain c_k:

$$\tilde{c}_m = \sum_{k=-\infty}^{\infty} \frac{1}{2\pi j} \oint_C \tilde{z}^{-k} z^{m-1} d\tilde{z}\, c_k \tag{3.17}$$

$$= \sum_k A_{mk}(\alpha)\, c_k \tag{3.18}$$

where $A_{mk}(\alpha)$ is the m-th row k-th column element of the warping matrix $\boldsymbol{A}_\alpha$ consisting of the warping factor α and the Cauchy integral formula yields [Pitz and Ney, 2005]:

$$A_{mk}(\alpha) = \frac{1}{2\pi j} \oint_C \tilde{z}^{-k} z^{m-1} d\tilde{z} \tag{3.19}$$

$$= \frac{1}{2\pi j} \oint_C \left(\frac{z-\alpha}{1-\alpha z}\right)^{-k} z^{m-1}\, d\tilde{z} \tag{3.20}$$

$$= \frac{1}{(k-1)!} \sum_{n=\max(0,k-m)}^{k} \binom{k}{n} \times \frac{(m+n-1)!}{(m+n-k)!} (-1)^n \alpha^{2n+m-k}. \tag{3.21}$$

The feature normalization can also be represented as a linear function that transforms the model parameters. This will be useful in more refined implementations of VTLN warping (to be discussed in detail in Chapter 5). A common representation of this linear function is the matrix transformation. The cepstral features are warped using the matrix representation as follows:

$$\boldsymbol{x}_\alpha = \boldsymbol{A}_\alpha \boldsymbol{x} \tag{3.22}$$

where $\boldsymbol{x}_\alpha = (\tilde{c}_1, \cdots, \tilde{c}_M)^\top$ and $\boldsymbol{x} = (c_1, \cdots, c_K)^\top$. α is the warping parameter and $\boldsymbol{A}_\alpha$ is the matrix transformation. The transform may also be directly applied to the dynamic features of the cepstra; the transformation matrix is block diagonal with repeating $\boldsymbol{A}_\alpha$ matrix.

$$\boldsymbol{x}_\alpha = \begin{bmatrix} \boldsymbol{A}_\alpha & 0 & 0 \\ 0 & \boldsymbol{A}_\alpha & 0 \\ 0 & 0 & \boldsymbol{A}_\alpha \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \Delta\boldsymbol{x} \\ \Delta^2\boldsymbol{x} \end{bmatrix} \tag{3.23}$$

where, $\boldsymbol{x}_\alpha$ is the warped cepstral coefficients, $\boldsymbol{x}$ is the static features, $\Delta\boldsymbol{x}$ and $\Delta^2\boldsymbol{x}$ are dynamic part of the cepstra.

The unwarped cepstral features could be multiplied with the linear transformation matrix to generate warped features. This results in significant computation savings since features need not be individually recomputed for each warping factor. Equation 3.21 can be represented in the form of a recursion [Tokuda et al., 1994b], which in turn, can be calculated as a transfor-

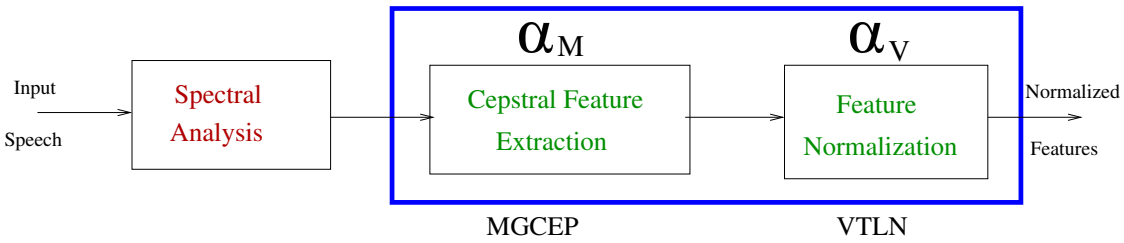


Figure 3.3 – Cascading the all-pass transforms in MGCEP features and VTLN transformation

mation matrix. This matrix representation of the MGCEP all-pass transform in the cepstral domain was presented by Saheer et al. [2010c] [3]:

$$\boldsymbol{A}_\alpha = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{(M-1)} \\ 0 & 1-\alpha^2 & 2\alpha(1-\alpha^2) & \cdots & (M-1)\alpha^{(M-2)}(1-\alpha^2) \\ 0 & -\alpha(1-\alpha^2) & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & (-1)^M(1-\alpha^2)\alpha^{(M-2)} & \cdots & \cdots & \cdots \end{bmatrix}$$

It can also be shown that the elements of this matrix can be estimated using the following recursive formula for $k > 1$ and $l > 1$

$$\boldsymbol{A}_\alpha(k,l) = \boldsymbol{A}_\alpha(k-1,l-1) + \alpha[\boldsymbol{A}_\alpha(k,l-1) - \boldsymbol{A}_\alpha(k-1,l)]$$

Estimating Warping Parameters

An all-pass transform based VTLN has been implemented in the MGCEP feature extraction with maximum likelihood (ML) optimization criterion. In the context of MGCEP features, VTLN can be considered as finding the optimal warping factor for each speaker. MGCEP already has an all-pass warping with $\alpha = 0.42$ approximating the mel-scale frequency warping for the sampling rate of 16kHz. Another stage of all-pass transform can be cascaded with the existing one to accommodate the VTLN warping as shown in Figure 3.3.

Consider the two all-pass transforms, one with warping factor α_V representing the VTLN frequency warping and other with warping factor α_M representing the warping in the MGCEP features. As shown by Acero [1993], the combination of two all-pass transforms is equivalent to an all-pass transform with single warping factor given by:

$$\alpha = \frac{\alpha_V + \alpha_M}{1 + \alpha_V \alpha_M} \tag{3.24}$$

The derivation can be found in Appendix B. The above equation can be used to estimate the

3. There was a slight error in the paper which has been since corrected.

different frequency warped MGCEP features, with $\alpha_M = 0.42$ for 16kHz speech and $|\alpha_V| \leq 0.1$. These features represent the grid of α values from which the best warping factor is to be estimated using the ML criterion.

VTLN can be used in both model training and target speech synthesis. During the model training, VTLN parameters are estimated for the training speakers and can be used to perform the speaker adaptive training. This will generate a better canonical HMM model for speech synthesis. During synthesis of speech for a target speaker, there are a few sentences spoken by the target speaker referred to as adaptation data. This adaptation data is used to estimate the warping factor for the target speaker. The warping factor can then be used to adapt the synthesized speech. This is performed in this chapter as a feature transformation after the speech parameters are generated from the synthesis model. This could be achieved by directly applying the transformation matrix or by using the recursion mentioned by Tokuda et al. [1994b] on the generated cepstra.

The steps involved in grid search based warping factor estimation for VTLN are as follows:

1. Generate warped MGCEP features with different warping factors.
2. Convert the HSMM model into an HMM model.
3. Align the warped features with the models and transcriptions to generate likelihoods.
4. Calculate the total likelihood for each speaker based on Equation 2.2 for each warping factor. The Jacobian of the transformation matrix (represented by $\log|A_\alpha|$) is used as the normalization constant [Uebel and Woodland, 1999].

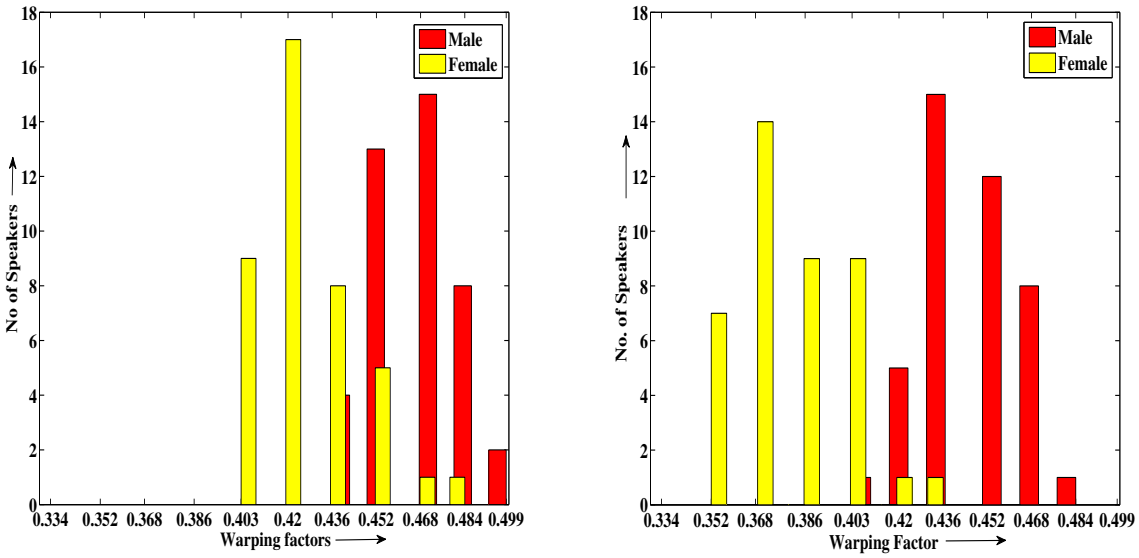
$$\hat{\alpha}_s = \underset{\alpha_s}{\operatorname{argmax}} \log|A_\alpha| p(\boldsymbol{x}_{\alpha_s} \mid \boldsymbol{\Theta}, w_s) \tag{3.25}$$

5. Find the best warping factor by comparing the likelihood scores.

Issues of dimensionality

HMM based speech synthesis systems require modeling of higher order features when compared to the speech recognition models. It was observed that the warping factor calculation was not successful with the higher (25^{th} or 39^{th}) order features, but worked with lower (12^{th}) order features. Similar observations can be seen in the literature [Emori and Shinoda, 2001, Hirohata et al., 2003]. The work of Hirohata et al. [2003] uses VTLN along with the MCEP (mel-cepstral) features in a similar way but restricting the estimation of the warping factor to using only the first few cepstral coefficients. The authors experimentally find that using only the first 4 coefficients of cepstral features gives better average voice in synthesis. However, the approach taken by Hirohata et al. [2003] is inaccurate due to the fact that the convergence of likelihood values is not guaranteed by warping the entire feature vector with the warping factors estimated from a few cepstral coefficients.

The failure of warping factor estimation for higher order features can be attributed to the presence of excitation harmonics, which could lead to a large likelihood mismatch even for a small warping. Furthermore, the estimation of optimal warping factor which alters the



(a) 25^{th} order system

(b) 25-12 system

Figure 3.4 – Warping factors estimated from 25^{th} order features. The 25-12 system initializes the features with the warping factors estimated from 12^{th} order features. Both graphs have same range for X-axis.

location of spectral peaks is disrupted by the presence of artifacts other than just the spectral peaks in the higher order features. It follows that the use of higher order features approaching 25^{th} or 39^{th} order MGCEP should be avoided when estimating warping factors. Instead, the warping factor estimated from the 12^{th} order features can be used as the seed values during the iterative VTLN training for higher dimensional features. It is observed that once a good initialization is given, the second iteration of VTLN training is able to estimate good warping factors even for higher order features. This phenomenon is illustrated in Figure 3.4. It can be observed from the figure that the distribution for warping factors estimated from the 25^{th} order has large overlap for male and female speakers with no proper separation of warping factors for female speakers. A more distinct bimodal distribution is observed when the warping factors are initialized with values estimated from the 12^{th} order features. More details on the challenges involved in VTLN implementation are presented in the following chapters.

Comparing with other linear transformations

CMLLR [Gales, 1998] is a powerful model based adaptation technique that can be shown to be equivalent to a feature transform [Pitz and Ney, 2005]. Details of CMLLR were presented in Chapter 1. CMLLR is a dominant adaptation approach due to the larger number of available parameters for speaker adaptation compared to VTLN. VTLN in combination with CMLLR has the potential to perform better, especially when there is little adaptation data or when using lower dimensional features for synthesis. CMLLR is used in the experiments presented below

as both a comparative baseline and also for its potential to be combined with VTLN.

3.3 Experiments & Results

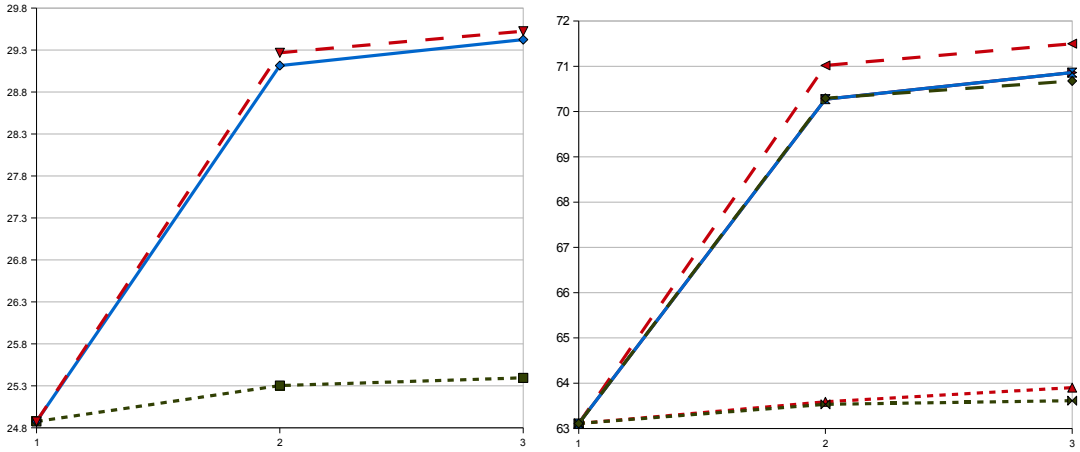
This chapter presents the initial implementation of VTLN for statistical parametric speech synthesis. An ML based grid search technique for VTLN using the MGCEP features is presented. This section presents the experimental evaluation of this technique using HMM based speech synthesis. The wall street journal (WSJ0 SI-84) database was used to build the speaker independent models. Evaluations were performed on the incremental speaker adaptive (S4-C3) data set of the WSJ Nov93 test specifications. In the training phase, warping factors were initially estimated using grid search for each speaker in the training data and the average voice models were iteratively trained by re-estimating the warping factors until convergence of the model likelihood on the training data. The same grid search technique was used to estimate the best warping factor for each test speaker using the available adaptation data from the corresponding speaker. The grid search for the warping factors was performed with $\alpha_M = 0.42$, and $-0.1 \leq \alpha_V \leq 0.1$ with a step size of 0.02. The two transforms were combined using the Equation 3.24.

Full context HSMM models were trained using the HTS 2.1 scripts by Yamagashi et al. [2007]. The feature vectors were divided into three streams. The 25 dimensional MGCEP features along with the corresponding Δ and Δ^2 parameters forms the first stream. The logF0 and its dynamics represented by the multi-state probabilistic distribution (MSD) forms the second stream. The five dimensional band aperiodicity and its dynamic features are the third stream in the feature vector. VTLN transforms were estimated on only the cepstral stream. HMM synthesis uses a longer context information (including approximately 40 contextual information) for labels and is termed as full-context labels.

The steps involved in the SAT model training using VTLN can be summarized as follows:

1. Build an average voice model using the un-warped cepstral features.
2. Use this average model and training transcriptions to compute the best warping factor for each speaker in the training set using the grid search approach as explained in section 3.2.
3. Iteratively re-estimate average voice models and VTLN warping factors until convergence of model likelihood on the training data (or possibly separate validation data) has been obtained.

The warping factor estimation was iterated twice to build improved average voice models. The adaptation sentences from each target speaker were used to estimate the optimal warping factor. This warping factor was used to transform the synthesized cepstral features using the recursion given by Tokuda et al. [1994b]. These warped cepstral features were then used to synthesize the speech of the target speaker. The results of VTLN transformation are compared with a standard adaptation technique (CMLLR). VTLN transformation were combined with the CMLLR transformation to enable additive improvements. Although VTLN cannot capture



(a) 12^{th} order features

(b) 25^{th} order features

Figure 3.5 – Log-likelihood scores during training. Dotted lines represent VTLN, dashed lines represent CMLLR and solid lines represent VTLN+CMLLR. 25^{th} order figure also includes the 25-12 case which has lower probabilities for VTLN and VTLN+CMLLR. X-axis denotes the increasing number of training iterations and Y-axis denotes the log-likelihood value.

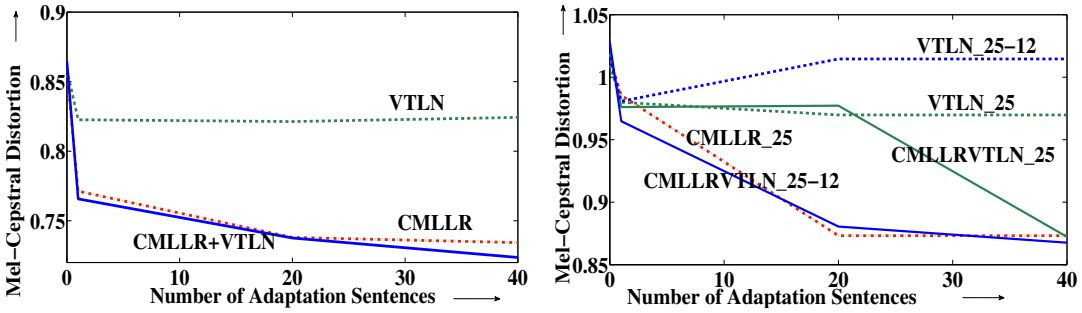
the entire characteristics of the speaker with the single warping factor, at least the gender characteristics can be accurately represented. Hence, VTLN has the potential to improve adaptation using little adaptation data along with other adaptation techniques like CMLLR.

Evaluation Metrics

The results can be evaluated objectively and subjectively. Objective evaluation of the synthesized speech is performed using a mel-cepstral distortion (MCD) measure, which is the average Euclidean distance between reference and synthesized mel-cepstral feature vectors. This can be considered to be equivalent to log-spectral distortion according to Parseval's theorem. The convergence of log-likelihood scores during training also presents a cue for the improvement in the average voice model training.

Subjective evaluation of the synthesized speech was conducted to determine mean opinion scores (MOS) for naturalness and speaker similarity. The naturalness was scored on a five point scale ranging from 1 to 5, where, 1 represents completely unnatural speech and 5 completely natural speech. Speaker similarity was also rated on a five point scale from 1 to 5, where 1 denotes speech from a totally different speaker and 5 denotes speech from exactly the same speaker.

Subjective evaluations were conducted on 60 randomly picked sentences from 10 different systems. 19 listeners were presented with the 60 sentences, randomly sorted to avoid any bias due to listening order. The 25^{th} order systems for VTLN, CMLLR and CMLLR combined with VTLN were tested with different amounts of adaptation data. These systems were also compared with their respective 25-12 counterparts, where the warping factors were initialized



(a) 12^{th} order features (b) 25^{th} order features

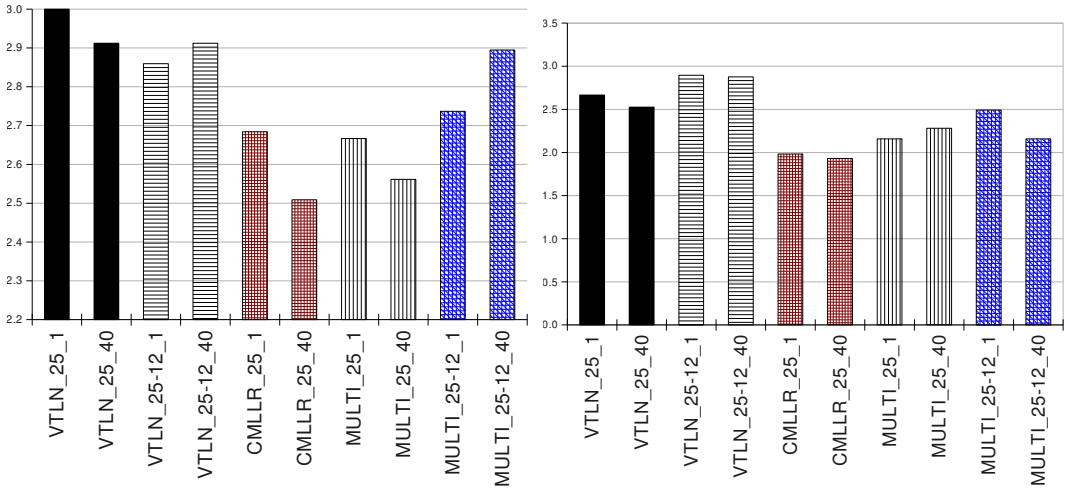
Figure 3.6 – Mel-Cepstral Distortion for synthesized speech. Dotted lines represent VTLN, dashed lines represent CMLLR and solid lines represent VTLN+CMLLR. 25^{th} order figure also includes the 25-12 case which has higher MCD for VTLN and lower MCD for VTLN+CMLLR. X-axis denotes the number of adaptation sentences starting from 0 (average voice) to a maximum of 40 sentences,Y-axis represents the MCD value.

from 12^{th} order and re-estimated using 25^{th} order cepstral features.

Results and Discussion

The experiments were performed on the MGCEP features with the analysis parameter, γ, equal to zero and with two different feature orders, 12 and 25. The results of objective evaluations are plotted as graphs. The log-likelihood scores increase with multiple iterations of each adaptation technique as shown in Figure 3.5. The MCD results for VTLN based feature adaptation are given in Figure 3.6. The feature order 25-12 represents the 25^{th} order features initialized with a warping factor estimated from 12^{th} order features. It can also be seen that CMLLR has additive improvements in performance in combination with VTLN. It can be seen that the average voice model trained with CMLLR and VTLN has better convergence during training and higher MCD during synthesis indicating that it should be a better average voice model. It can be observed for 25^{th} order features that VTLN and CMLLR combined with VTLN have lower MCD than CMLLR when only a single adaptation sentence is available. Also, the adapted speech with VTLN in combination with CMLLR gives lower MCD for any amount of adaptation data, suggesting that VTLN can contribute to improvement of the synthesized speech.

Results for subjective evaluations are shown in Figure 3.7, which shows MOS for naturalness and speaker similarity. Subjective tests were conducted on 10 different systems. These include VTLN, CMLLR and CMLLR+VTLN for 25^{th} order and 25_12 systems with adaptation using 1 and 40 sentences. It is observed that VTLN systems are preferred over other systems for the naturalness cue. Also, VTLN combined with CMLLR is preferred as having better similarity to the voice of the original speaker when compared to CMLLR. The subjective evaluations as such only have limited statistical significance since it is observed that CMLLR system was not preferred at all for naturalness or speaker similarity. But, these scores support the results from objective evaluations emphasizing the fact that VTLN can perform additive improvements to



(a) MOS for speaker similarity

(b) MOS for naturalness

Figure 3.7 – MOS for naturalness and speaker similarity of synthesized speech. The systems are named as Adaptation-Type_Feature-order_Number-of-Adaptation-Sentences. For example, VTLN_25_1 represents the VTLN adaptation system for 25^{th} order with 1 adaptation sentence. The 25-12 system is the 25^{th} order system with warping factors initialized with the values estimated from 12^{th} order system. MULTI represents the combination of VTLN and CMLLR adaptation techniques.

CMLLR. The Table 3.1 shows the paired Wilicoxon signed rank test result for demonstrating the differences between different pairs of systems.

3.4 Summary of Contributions

This research has successfully implemented VTLN based adaptation for statistical speech synthesis and incorporated the warping at the feature extraction stage of MGCEP features. It was observed that the VTLN parameters can be accurately estimated from much less adaptation data, as little as a single sentence. VTLN adaptation can estimate the correct gender charac-

Label Correspondence

#	Figure_Label
A	VTLN_25_1
B	VTLN_25_40
C	VTLN_25-12_1
D	VTLN_25-12_40
E	CMLLR_25_1
F	CMLLR_25_1
G	MULTI_25_1
H	MULTI_25_1
I	MULTI_25-12_1
J	MULTI_25-12_1

Significance at 5% level for **SpeakerSimilarity**

	A	B	C	D	E	F	G	H	I	J
A	-	0	0	0	1	1	1	1	1	0
B	0	-	0	0	0	1	0	1	0	0
C	0	0	-	0	0	1	0	1	0	0
D	0	0	0	-	0	1	1	1	0	0
E	1	0	0	0	-	0	0	0	0	0
F	1	1	1	1	0	-	0	0	0	1
G	1	0	0	1	0	0	-	0	0	0
H	1	1	1	1	0	0	0	-	0	1
I	1	0	0	0	0	0	0	0	-	0
J	0	0	0	0	0	1	0	1	0	-

Significance at 5% level for **naturalness**

	A	B	C	D	E	F	G	H	I	J
A	-	0	0	0	1	1	1	1	0	1
B	0	-	1	1	1	1	1	0	0	1
C	0	1	-	0	1	1	1	1	1	1
D	0	1	0	-	1	1	1	1	1	1
E	1	1	1	1	-	0	0	1	1	0
F	1	1	1	1	0	-	1	1	1	0
G	1	1	1	1	0	1	-	0	1	0
H	1	0	1	1	1	1	0	-	0	0
I	0	0	1	1	1	1	1	0	-	1
J	1	1	1	1	0	0	0	0	1	-

Table 3.1 – First table shows the correspondence between the labels in the Figure 3.7 and the system names in the rest of the tables. The second and third tables indicate significant differences between pairs of systems, based on Wilcoxon signed rank tests with alpha Bonferoni correction (5% level); '1' indicates a significant difference.

teristics of the speech with a single adaptation sentence, and hence the adapted sentence sounds more similar to the original speaker. The warping factor estimation for higher order features can be improved by initializing with values estimated from lower order features. It was also observed that VTLN gives additional improvements with CMLLR adaptation. The work presented in this chapter was initially presented in [Saheer et al., 2010c].

4 Theory and Practice: A unified view

Some of the findings in the process of trying to make VTLN work for statistical parametric speech synthesis have provided additional insight into ASR and enabled an overall more consistent view of VTLN theory and it's application. There are a number of challenges involved in the implementation of VTLN for ASR and TTS. Although not directly obvious, a few publications can be found in the literature dealing with the challenges in implementing VTLN for ASR. There has been very limited investigation, if any, in the area of speech synthesis. This chapter reviews the challenges in ASR and presents the new challenges faced in the VTLN implementation for statistical parametric speech synthesis. This chapter also presents a number of techniques to work around these challenges and improve the performance of VTLN warping factor estimation especially for the HMM-based TTS system.

4.1 Related Work

The challenges in implementing VTLN can be attributed to the fact that VTLN represents a spectral transformation. To this end, the implementation of VTLN for speech recognition involves reconstruction of the untruncated spectrum from the cepstral features that demands some kind of interpolation for feature space transformations. Apart from the challenges in ASR, VTLN for TTS has to tackle some additional challenges. There are no prior studies that can be found in the literature with respect to VTLN for parametric synthesis. Hence, this work is the pioneer in this area. Challenges presented for ASR are also applicable for TTS. All the additional challenges in TTS might also be present in ASR. But, they become prominent only in the case of TTS due to the higher feature dimensionality.

This section summarizes two major challenges in using VTLN for ASR. VTLN transformation can either be carried out in the spectral-domain as a part of the feature extraction step or in the cepstral-domain since the linear operations in the spectrum have their equivalence in the cepstral domain. Unfortunately this equivalence is only an approximation since intermediate steps (for instance, filter-bank smoothing) in cepstrum computation results in truncation and loss of information. This is the first challenge presented in this section which deals with the

spectral reconstruction problem due to the removal of information. The second challenge is using the Jacobian term to normalize the likelihood scores for VTLN warping factor estimation. This factor appears to hinder the accurate estimation of the warping factors especially for a target speaker. The first challenge is unique to ASR, while the second one proves to be a problem for statistical parametric speech synthesis as well.

Spectral reconstruction problem

Speech recognition systems favour features that are invariant to undesirable information sources, such as the speaker, pitch harmonics, etc. To this end, most of the commonly used features in ASR use a bank of filters to smooth off any spectral variations across utterances from different speakers. The commonly used ASR features are mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) features. Both of these feature extraction techniques involve a bank of filters which also performs frequency warping (mel or bark scale) that matches the human perception.

One of the first implementations of VTLN for MFCC features was by Lee and Rose [1996]. The VTLN frequency warping was embedded into the bank of filters along with the mel scale warping. A grid of warped cepstral features were generated for each speaker and a grid search technique was used to estimate the optimal warping factor leading to high space and time complexity for the warping factor estimation. The Jacobian term was ignored. It was shown by Pitz and Ney [2005] that the linear transformations on the cepstral features are equivalent to the linear transformations in the spectral domain. Kim et al. [2004] designed the VTLN transformation as a linear transformation of the cepstral features thus, drastically improving the time complexity involved in the VTLN transform estimation. Though, not a straightforward technique with the MFCC and the PLP features, there were implementations that embedded the VTLN transformation matrix into the discrete cosine transformation (DCT) [Umesh et al., 2005, Panchapagesan and Alwan, 2009] step of the feature extraction. There were even further improvements to this technique that lead to show that the same performance can be obtained with linear transformation of cepstra as with the traditional grid search technique [Sanand and Umesh, 2012].

Warped cepstral features can be estimated by linear transformation of the cepstral features instead of recomputing the features for each warping factor. This method was further improved upon by exploiting the equivalence of model and feature transformations and thus, estimating the VTLN warping factor from the model parameters using efficient expectation maximization (EM) algorithms [Panchapagesan and Alwan, 2009, McDonough, 2000, Umesh et al., 2005].

Jacobian Normalization

Jacobian normalization should be applied to the likelihood score calculations following a feature transformation (details of the derivation presented in section 4.3). In spite of this, it

has been observed in ASR studies that use of Jacobian normalization tends to have an adverse effect on the performance of VTLN especially during testing. It can be seen in the literature that most VTLN implementations either ignore the Jacobian normalization or replace it with cepstral mean/variance normalization [Uebel and Woodland, 1999, Lee and Rose, 1998, Garau, 2008]. One recent study by Sanand et al. [2009] on mismatched train and test conditions addressed this issue by compensating with a variance adaptation on top of VTLN along with Jacobian normalization. This approach is effectively able to further compensate the change in feature variability caused by the VTL transform, which is otherwise compensated only by Jacobian normalization. In reality, adaptation of the variance effectively obviates the need for Jacobian normalisation since this provides a kind of score normalisation. The performance improvement was observed in both matched and mismatched test conditions. A systematic study of Jacobian normalization was performed by Pitz [2005].It was shown that though the Jacobian normalization affected the warping factor values, it did not have much influence on the recognition performance. A few techniques like scaling the Jacobian factor were also proposed to reduce the influence of the Jacobian normalization. Finally, the omission of the Jacobian determinant was justified. It was mentioned that VTLN does not fail without proper normalization due to the limitation of the warping factors and the specification of the warping function. The resulting transformation was claimed not to have enough degrees of freedom and the matrix is kept close to identity.

4.2 Additional Challenges in TTS

In this section, the main issues for the successful application of VTLN for statistical parametric speech synthesis are analyzed. In particular, the challenges that are distinct from those encountered in past work concerning VTLN, (most notably distinct from its application to ASR) are analyzed in detail. These challenges centre around two main factors, numerical / computational and subjective / perceptual.

First of all, the impact of numerical modelling factors, particularly the impact of feature extraction for synthesis on vocal tract length estimation is analyzed. It is evident that ASR and TTS use feature analysis methods that differ quite markedly and this can have an impact on VTLN. In particular, the high feature dimensionality can pose significant difficulties (as was shown in Chapter 3). While the truncation of the cepstrum and the truncation of the transformation matrix can add additional challenges in the calculation of inverse transformation during synthesis. Thus, the challenges that TTS features introduce to VTLN are examined. Most of these challenges also exists in ASR but become more prominent in TTS due to the higher feature dimensionality.

Second factor is that evaluation of TTS is conducted through subjective testing, hence, understanding of human perception and design of appropriate tests are necessary. In this chapter, some preliminary analysis on perception of vocal tract length vis-à-vis speaker similarity and whether or not there exist any correlates with low-level acoustic features are presented. Some

thoughts on appropriate means to evaluate VTLN as a speaker adaptation / transformation approach in statistical parametric speech synthesis are also presented.

4.2.1 Modelling factors

The main differences between VTLN for ASR and TTS are in the feature type and feature dimensionality. Usually, MFCC features of the order of 12 are used in speech recognition. This provides a coarse representation of the spectral envelope and in turn the formant structure in the speech. The MGCEP features used for statistical speech synthesis have very high dimensionality — of the order of 25 or 39 (for 16kHz sampling rate and even higher for higher sampling rates of the speech) — when compared to ASR features in order to obtain good synthesis quality by maintaining the fine structure of the spectrum.

Higher order cepstral coefficients will capture aspects of spectral fine structure which may cause problems when estimating the values of α. In particular VTL is normally considered as being related to formant location which in turn is captured sufficiently by lower order cepstra. Despite this, the higher order cepstra can still have a significant (and potentially detrimental) impact on warping factor estimation since they will make a significant contribution to the likelihood score calculation.

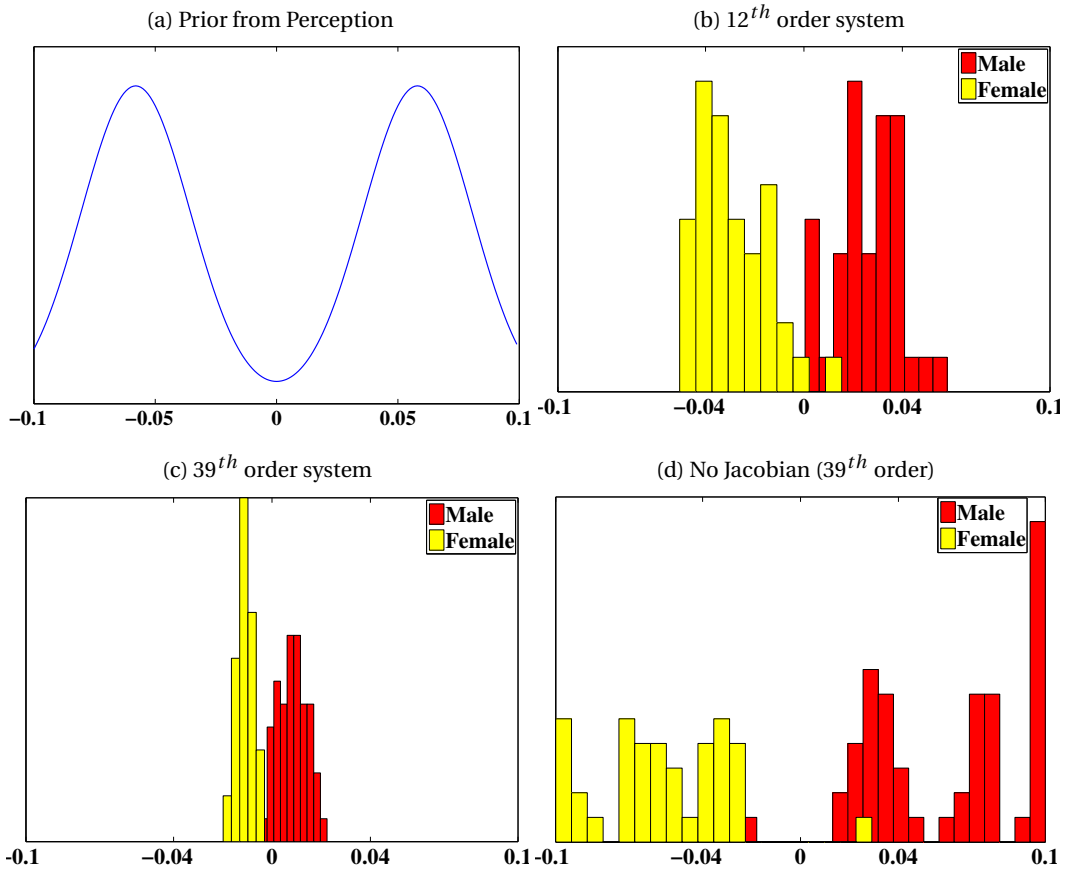
As an example of this phenomenon, it can be observed from the Figure 4.1c that the range of warping factors estimated from high order feature analysis is extremely narrow compared to those (shown in Figure 4.1b) for a more typical feature analysis order in ASR. In reality, such a narrow range of warping factors will be imperceptible in most cases. By contrast, it can be seen from Figure 4.1d that the omission of Jacobian normalization (to be discussed further below) will also result in warping factors far from what is expected to be appropriate, with divergence towards the boundaries.

Various means are discussed in the next section to pragmatically address these issues as they relate to warping factor estimation for TTS.

Jacobian Normalization

It was observed that the Jacobian normalization has an important role in VTLN for statistical parametric speech synthesis, in particular, due to the use of high order feature analysis [Saheer et al., 2010c]. One effect of Jacobian normalization is to penalize severe warping factors, thereby reducing the range of α that would otherwise be estimated in its absence. This is necessary since the warping transformation itself results in narrower variance of the transformed features that would result in an unfair bias towards extreme warping factors. This effect can be visualized in Figure 4.2 for different feature orders. It is evident that the Jacobian normalization term becomes more significant as the feature order increases and cannot be ignored as has been done in previous ASR studies.

Figure 4.1 – Distributions over warping factor value. The abscissa is α, the warping factor. Ordinate represents the frequency of α.



It is apparent that the inclusion or omission of Jacobian normalization does not give desirable warping factors, hence, additional measures are required, as presented in the next section.

4.2.2 Subjective and perceptual factors

The major concern in this work is the use of probabilistic techniques for the application of VTLN to TTS; the ultimate goal is to find the best methods based on subjective evaluation. At times, this may require divergence from strict adherence to the theory. In short, subjective evaluation is the primary determinant of the success of the TTS approaches presented in this work, and as such one needs to understand perception of VTLN in statistical parametric synthesis. There are objective metrics that are easy to evaluate, but, usually not reliable and meaningful. The subjective metrics by contrast, are expensive in time and effort and may even provide misleading results if not properly designed. Hence, understanding perception of VTLN is critical to measuring impact of its application.

Perceptual experiments are the only way to evaluate the improvements from new techniques in

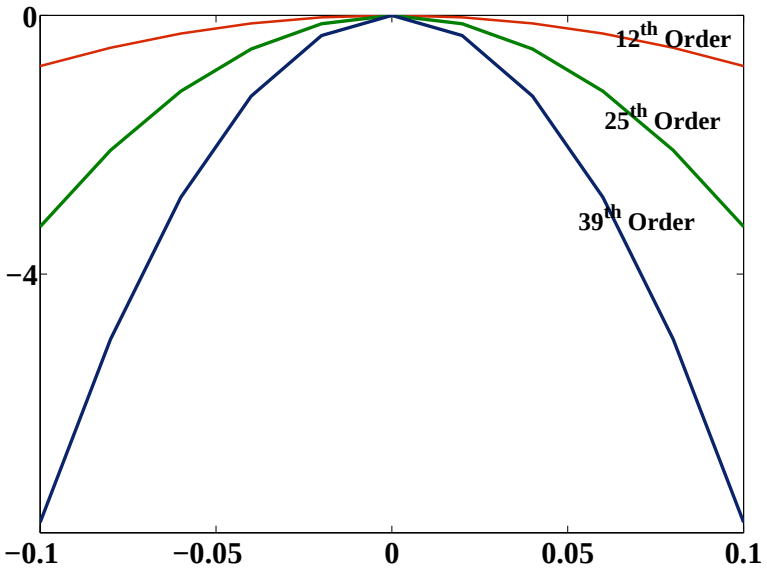


Figure 4.2 – Jacobian calculated as $\log|\boldsymbol{A}|$ for various feature dimensions. The abscissa is α, the warping factor and the ordinate is value of the Jacobian determinant.

TTS. Evaluating VTLN is a non-trivial task since very few speaker characteristics are captured. This also poses a challenge of identifying speakers who can demonstrate the performance improvements with VTLN. Hence, this work also addresses the issue of finding the best speakers for evaluating VTLN.

4.3 Proposed Solutions

There are very limited studies in the literature that address any solutions to the above challenges. The implementation of VTLN for parametric synthesis as such is a novel area of research and so are the challenges in using VTLN for synthesis. This section proposes a number of solutions to get around these challenges. These are evaluated using TTS or ASR experiments as shown in the next section.

Derivation for Jacobian Normalization

VTLN needs a Jacobian normalization term to compare the likelihood values for different warping factors. This term comes into effect for any feature adaptation technique which can be shown to be equivalent to a model adaptation technique. The following derivation demonstrates how this term shows up in a feature adaptation formulation.

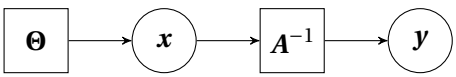


Figure 4.3 – Generative model for vocal tract "warping"

Assume a model, $\boldsymbol{\Theta}$, generates a sample, $\boldsymbol{x}$ as shown in Figure 4.3. The sample is then distorted

by a linear transform, $\boldsymbol{A}^{-1}$, a function of α, to give an observation $\boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{x}$. Here, we follow the convention that $\boldsymbol{A}$ is a feature transform so the generative transform is $\boldsymbol{A}^{-1}$. The goal is to find an optimal value, $\hat{\alpha}$, of α. Bayes's theorem gives the maximum *a posteriori* estimator:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, p_{\alpha}\left(\alpha \mid \boldsymbol{y}, \boldsymbol{\Theta}\right) \propto\ p_{\boldsymbol{y}}\left(\boldsymbol{y} \mid \alpha, \boldsymbol{\Theta}\right) p_{\alpha}\left(\alpha \mid \boldsymbol{\Theta}\right). \tag{4.1}$$

To evaluate the first term on the RHS of equation 4.1, notice that the model generates $\boldsymbol{x}$ rather than $\boldsymbol{y}$, so it needs a change of variable $\boldsymbol{y} \rightarrow \boldsymbol{x}$. The Jacobian determinant for the change of variable is,

$$J = |\boldsymbol{A}|, \tag{4.2}$$

where the notation is taken to mean the determinant of the matrix. So,

$$p_{\boldsymbol{y}}\left(\boldsymbol{y} \mid \alpha, \boldsymbol{\Theta}\right) = |\boldsymbol{A}|\, p_{\boldsymbol{x}}\left(\boldsymbol{A}\boldsymbol{y} \mid \alpha, \boldsymbol{\Theta}\right). \tag{4.3}$$

The second term on the RHS of equation 4.1 is a prior on α. Notice that α is actually independent of the model, $\boldsymbol{\Theta}$, so it could be written unconditional. However, α is posterior to the training data, $\boldsymbol{D}$, that was used to train $\boldsymbol{\Theta}$. So, equation 4.1 can be evaluated as

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} |\boldsymbol{A}|\, p_{\boldsymbol{x}}\left(\boldsymbol{A}\boldsymbol{y} \mid \alpha, \boldsymbol{\Theta}\right) p_{\alpha}\left(\alpha \mid \boldsymbol{D}\right). \tag{4.4}$$

Notice that the prior is not normally considered in most implementations of VTLN.

Use of a prior distribution of warping factors

The MAP estimation of warping factors using the Bayesian framework is given by:

$$\hat{\alpha}_s = \underset{\alpha_s}{\operatorname{argmax}} |\boldsymbol{A}|\, p(\boldsymbol{x}_{\alpha_s} \mid \boldsymbol{\Theta}, w_s)\, p(\alpha|\boldsymbol{\Theta}) \tag{4.5}$$

The variables are same as in Equation 2.2 with the addition of a Jacobian normalization term ($|\boldsymbol{A}|$) and the prior probability distribution for the warping factor α, denoted by $p(\alpha|\boldsymbol{\Theta})$. The prior on the warping factor (in the above equation) has been omitted from the VTLN implementations presented in the literature. Ignoring a prior normally corresponds to assuming a flat prior. Where sufficient data is available, this is often a reasonable approach. Conversely, when little data is available use of a prior can be important. We can assume that the prior on the warping factor should not be flat, more precisely:

– It should tend to zero at the extreme values ± 1.
– It should be bimodal, representing the distribution of male and female population.

Objectively, the prior can be measured via a histogram of warping factors calculated over a large number of speakers, for each of whom a large amount of data exists. Such histograms are shown in Figure 4.1b, and moments can be measured to infer a parametric distribution. Here, we use a two-component beta mixture, transformed to span the range ±1:

$$p_{\propto}(\alpha \mid \boldsymbol{D}) \sum_{g \in \{m,f\}} (1+\alpha)^{p_g - 1}(1-\alpha)^{q_g - 1}, \tag{4.6}$$

where $\{p_m, q_m\}$ and $\{p_f, q_f\}$ are the pairs of beta parameters for male and female speech respectively, as in Figure 4.1a.

Note that omitting the Jacobian determinant from the likelihood calculation has the effect of using a prior with a PDF proportional to the inverse of the Jacobian (c.f. Figure 4.2). It biases α away from zero, enhancing the relative separation of the male and female modes. It can be seen from Figure 4.4a that the prior does not have much impact on warping factor estimation for the training data due to availability of sufficient data. The changes are expected to be seen only in the warping factors for the test data. This in turn can explain why it has been observed by earlier researchers that omitting Jacobian normalization improves performance especially in testing: during testing there is often insufficient data to generate a reliable estimate of warping factor, thus means to increase the spread of warping factors by omitting the Jacobian term acts as a reasonable prior in the case of low dimensional feature analysis.

Using likelihood scaling

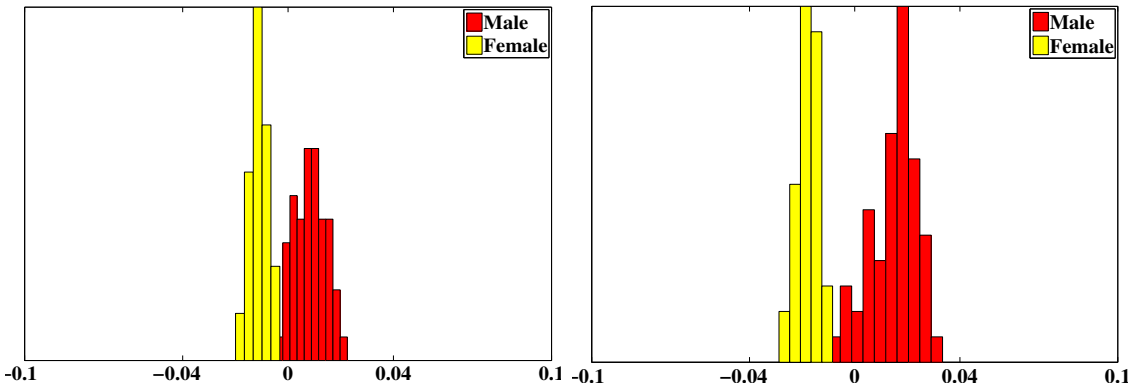
In large vocabulary ASR, it is common to use a language model scale factor that in fact compensates for under estimation of acoustic likelihoods. This in turn is necessary because successive acoustic frames have much higher correlation than is captured by the HMM. Applied more correctly to the acoustic calculation, we might expect that the likelihood correction should apply to the likelihoods, but not to the Jacobian. In fact, this was investigated by Pitz [2005], who applied the factor to the Jacobian analogous to the language model scale. This suggests an estimator of the form

$$\hat{\alpha}_s = \underset{\alpha_s}{\operatorname{argmax}} |\boldsymbol{A}| \left[p(\boldsymbol{x}_{\alpha_s} \mid \boldsymbol{\Theta}, w_s) \right]^{\Psi} p(\alpha|\boldsymbol{\Theta}) \tag{4.7}$$

where, Ψ represents the scale factor for boosting likelihoods. The effect of the scale factor value "2" is shown in Figure 4.4b. The "optimal" scale factor is estimated empirically.

Using lower order features

VTLN is intended to transform the location of spectral peak positions, thus logically its estimation should be based upon only the coarse spectral envelope. To illustrate this, reconstructed spectra from different MGCEP analysis orders are plotted in Figure 4.5. It can be noted that



(a) Beta Prior Distribution (b) Scale factor

Figure 4.4 – Warping factors estimated from 39^{th} order features with a scale factor of 2 for the likelihoods and with a beta prior distribution. x-axis represents α values and y-axis represents the frequency of α values.

the cepstral order of 9 or more represents at least the first two formants. As the cepstral order increases beyond this, the finer details of the spectrum are better represented and the pitch harmonics also creep into the spectrum. We suppose that the spectral envelope as it relates to VTL is most accurately represented by an order around 12. This is consistent with use of feature order of 12 in ASR. In statistical parametric TTS, one means to overcome the problems posed by higher order cepstra in VTLN implementation is to estimate VTLN warping factors using only a subset of the cepstral features, specifically the first 13 (including C0). While this means that the implementation of VTLN is no longer guaranteed to increase likelihood on the entire observation space, this work is more concerned with the implications of such an approach on subjective evaluations as this is of primary importance for TTS. As mentioned earlier, similar approaches are also found in the work of [Hirohata et al., 2003] which restricts the estimation of the warping factors from only the first few cepstral coefficients. The VTLN parameter estimated from the first 4 coefficients of cepstral features was shown to generate better average voice in synthesis. However, as mentioned above the approach taken by Hirohata et al. [2003] is inaccurate due to the fact that the convergence of likelihood values is not guaranteed by warping the entire feature vector with the warping factors estimated from a few cepstral coefficients. Additionally, as seen from Figure 4.5, a feature dimensionality of 4 does not necessarily even represent the spectral peaks.

Approximation for inverse of the transformation matrix

Truncation of the cepstra and hence, truncation of the transformation matrix generates some inconsistencies in the calculation of the inverse of the transformation matrix. This inconsistency can be attributed to the fact that the matrix A_{α}^{-1} does not represent an exact bilinear transformation matrix and hence, may not perform the desired warping.

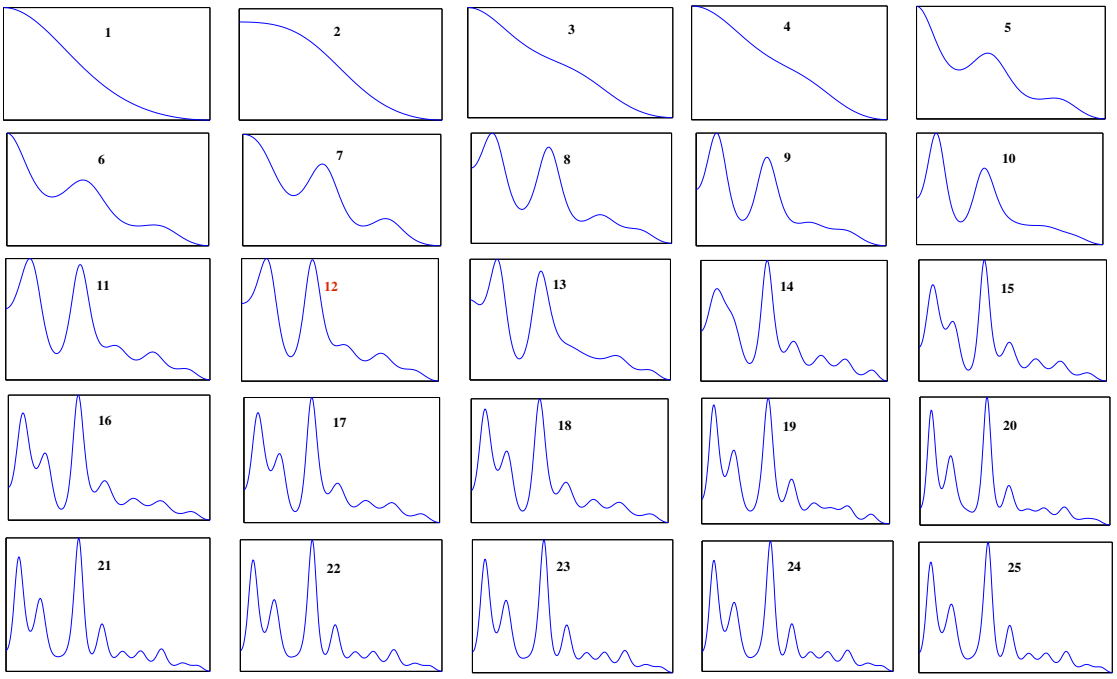


Figure 4.5 – Spectra reconstructed with cepstral features of the order of 1 to 25. The abscissa represents frequency and ordinate represents the spectral power.

Fortunately, the inverse of the bilinear transformation can be approximated as:

$$\boldsymbol{A}_{\alpha}^{-1} \approx \boldsymbol{A}_{-\alpha} \tag{4.8}$$

However, it should be noted that the expression is indeed an approximation (one just provides a more numerically stable way of computing the other). The matrices A_{α}^{-1} and $A_{-\alpha}$ are two possible approximations of the exact infinite transformation. While the former is more consistent with the estimation procedure, the latter is closer to an actual bilinear VTL warping, especially for lower order cepstral components.

This approximation can be justified by the fact that Equation 4.8 is exact in the infinite untruncated case and that it is used only during synthesis. This approximation does not have any effect on modelling or warping factor estimation. Only a determinant of the transformation matrix needs to be calculated during the warping factor estimation

Approximation of the log determinant

The matrix for expressing VTLN as a linear transformation of the cepstrum can be “large”, especially when using higher order cepstral features. As the order of the cepstra and the value of the warping factor increase, the calculation of the determinant becomes numerically unstable. The logarithm of the determinant can be calculated accurately as the sum of the

logarithms of the eigenvalues of the matrix thus:

$$\log|\boldsymbol{A}| = \frac{1}{2}\sum_{i=1}^{N}\log(e_i e_i^*) \tag{4.9}$$

where, e_i represents an eigen value of the matrix $\boldsymbol{A}$ and e_i^* represents the conjugate of the complex number e_i.

Perceptual evaluation of VTLN

The solution for the perceptual problems with VTLN is to perform some experiments to evaluate VTLN using perceptual experiments. Section 4.4.1 provides details on perceptual experiments that were carried out to establish a "subjective ground truth" for perception of VTL warping. These experiments also provide some insight to acoustic correlates of VTL (in particular pitch) and provide a basis for establishing a prior distribution of warping factors (as discussed in Section 4.3). The appropriate means to evaluate VTLN as speaker adaptation method, in particular, given its limitations with respect to reproducing specific voice characteristics are also discussed.

It is a non-trivial task to select speakers who can demonstrate the subtle differences when using different techniques for VTLN in TTS. Subjective evaluations are more tedious to perform and, hence, it can be postulated that objective scores can shed some light on the performance differences between different VTLN implementations. In such a case, subjective evaluations can be limited to a few selected speakers or systems. To this end, experiments are performed to find any correlations (if exists) between subjective and objective evaluations for different VTLN implementations in synthesis. This will in turn help in selecting the best speakers who can demonstrate the performance of VTLN. The details of the experiments and results is given in section 4.4.1.

4.4 Experiment & Results

A number of challenges and their corresponding solutions were proposed in the previous sections. The solutions proposed need to be evaluated experimentally to check if the theory can be matched to the practice. This section in turn presents a few recognition and synthesis experiments to evaluate each of the proposed solutions and in turn, suggest some perceptual evaluations that aid in effective evaluation of VTLN performance for statistical parametric speech synthesis. Following is the list of diverse evaluations presented in the following sections and corresponding motivations for each experiment.

– Estimation of warping factors with the static and dynamic parts of the cepstra in order to find the best set of feature representations for VTLN.
– Recognition experiments to evaluate the different solutions in this chapter and also as a unified framework for ASR and TTS.

– Experiments for perceptual impact of warping factors that demonstrate the range of warping that can be perceived and preferred by listeners and hence ideal for VTLN. These experiments should also provide a prior probability distribution for warping factors.
– Speaker selection experiments for VTLN evaluation using objective and subjective measures. These experiments will help select the best speakers who can represent the performance differences with VTLN adaptation.

4.4.1 Evaluating the proposed solutions to modelling problems

The following experiments used the WSJCAM0 British English read speech database. 39^{th} order acoustic models for TTS were trained using hidden semi-Markov models (HSMM) with only a single Gaussian PDF per state. The bilinear transform based VTLN was applied on the mel-cepstral (MCEP) features with a warping factor within the range of -0.1 to 0.1.

Separating the static and dynamic cepstra for VTLN

As the theory suggests, Jacobian normalization should be used for warping factor estimation. When the feature stream contains dynamic components, the transformation can be expressed as follows.

$$\hat{c} = \begin{bmatrix} \boldsymbol{A} & 0 & 0 \\ 0 & \boldsymbol{A} & 0 \\ 0 & 0 & \boldsymbol{A} \end{bmatrix} \begin{bmatrix} c \\ \Delta c \\ \Delta^2 c \end{bmatrix}, \tag{4.10}$$

where $\boldsymbol{A}$ is the transformation on the static features and can be directly applied to the dynamic part of the cepstra as well. In this section, it is shown experimentally that the warping factors estimated from the static features are more accurate. Estimating warping factors as a transformation on the cepstrum should take into account the fact that the feature stream usually contains dynamic features which can disrupt the warping factor estimation. Table 4.1 shows the warping factors estimated using static and dynamic feature vectors separately for a male and female speaker.

Gender	**Static**	Δ	Δ^2
Male	0.0195	0.0100	-0.0145
Female	-0.0260	-0.0142	0.0134

Table 4.1 – Warping factors for components of feature vectors

Ideally, the derivatives should not affect the warping factor estimation and the static and dynamic parts of the cepstra should yield the same value for the warping factors. The reason why the dynamic features are not able to estimate exactly the same warping factors as the static features is because they no longer represent the spectral envelope and hence, cannot accurately estimate the warping factor. This problem will not be observed in VTLN techniques

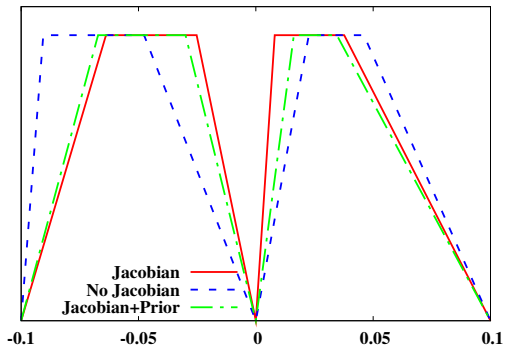


Figure 4.6 – Warping factors estimated for test (Nov93 Eval) data from 13^{th} order MCEP features with and without Jacobian normalization and prior. abscissa represents the α values and the ordinate represents the frequency of α values.

embedded into the feature extraction step (like warping the filter banks of MFCC features) which estimate dynamic features from the already warped static features.

Recognition performance

Speech recognition experiments are presented here showing that Jacobian normalization should be used in VTLN. The hidden Markov models were built with 13 dimensional cepstral features with Δ and Δ^2 for the (US English) WSJ0 database. The models were built using single component mixture PDFs to demonstrate the maximum impact of VTLN, and because only single component mixture models can be used in synthesis. In particular, this was to avoid the situation where multi-mixture models either over-fitted, or modeled the speaker variabilities that could be attributed to VTLN. Similar observations can be seen in Welling et al. [1999] where single component mixture models are used for VTLN parameter modeling. The performance differences are not statistically significant, but support the fact that Jacobian normalization should not degrade the performance. These results are similar to ones shown by Pitz [2005], where experiments are performed using a scale factor for the Jacobian analogous to inverse scaling of the likelihood.

Moments of symmetric beta prior distributions were estimated from the warping factors for the conversational speech database presented by Garau [2008]. The warping factors for the test speakers are shown in Figure 4.6, which shows that using a prior distribution can estimate similar warping factors to those that might result from not using Jacobian normalization. The recognition performance is not significantly affected by the separation of the warping factors. However, the separation is important in statistical speech synthesis, which demands distinct warping factors for each speaker to bring as many characteristics of the speaker as possible in the synthesized speech. A scaling factor may be needed in TTS where the prior is insignificant when combined with the large likelihood scores produced by the higher order features.

SI-model	VTLN		
	No Jacobian	Jacobian	Jacobian+Prior
22.16	19.43	19.33	19.49

Table 4.2 – WER for 13^{th} order features on the Nov93 Eval (hub task h2_p0)

This section takes various techniques proposed for application of VTLN to ASR (prior distribution, or omission of Jacobian) and compares them. This confirms some of the previous research and helps shed some light on some of the more empirical work (e.g. omission of the Jacobian is like using a poor prior.) A secondary objective here is to provide a unified approach to VTLN (from ASR & TTS perspective). Ideally the conclusions from this work should be consistent for ASR and TTS.

Perceptual impact of warping factors

In order to evaluate VTLN for speech synthesis it is important to understand perception of VTL transformations, especially with respect to speaker similarity. In studying perception of VTL, the data collected can be used as the basis for constructing valid prior distributions for the warping parameter, α. This prior distribution can then have direct application in the MAP solution to the warping factor estimation.

VTL varies across speakers resulting in corresponding changes in the spectral peak positions. Conversely, warping the spectral frequencies of a recording should bring in approximately the same variation that is audible due to the differences in vocal tract length. A preliminary experiment conducted on a speaker's voice using analysis-synthesis with different levels of warping provides evidence for this fact. It was noted that whenever the spectral frequencies are expanded, the speech sounded more "feminine" as if from a shorter vocal tract. Also, whenever the spectral frequencies are compressed, the speech sounded more "masculine" as if from a longer vocal tract. Both phenomena are observed in spite of using the original pitch of the speaker. These observations led to the design of a subjective evaluation to determine the perceived warping factors for a set of speakers. The values obtained from these evaluations are compared with the warping factors derived from the model.

The HMM speech synthesis system (HTS) by Yamagishi et al. [2009b] was used to build average voice models using 39^{th} order cepstral features along with Δ and Δ^2 values of MGCEP features. Experiments were performed on the WSJCAM0 (British English) database with 92 speakers in the training set. The details of the synthesis system can be seen in Chapter 1. Warped sentences were directly synthesized using the negative warping factors (instead of using default $\alpha = 0.42$), which is equivalent to applying the VTLN matrix $A_{(-\alpha)}$ to the untruncated cepstra.

Pitch and vocal tract length ideally should not be treated as independent parameters, although techniques have been proposed to implement VTLN based on pitch [Faria and Gelbart, 2005]. It is not in the scope of this work to implement such a system, but this factor is taken into

Table 4.3 – Frequency of female speakers with different combinations of vocal tract length and pitch

Pitch Vs "α" group	**1**	**2**	**3**	**4**	**5**	**6**
Low (159-190)	1	5	4	1	0	0
Medium (191-222)	0	4	8	8	1	1
High (223-255)	0	1	1	4	0	1

consideration in the design of perceptual experiments; more specifically, the speakers selected in this experiment cover different combinations of pitch and VTL.

Experiments were performed only on the female speakers of the WSJCAM0 database. There were a total of 40 female speakers in the training set. A subset of 20 female speakers were selected in such a way that they cover the different possible combinations of pitch and VTL. The gender restriction helps to minimize the size of the evaluations. The distribution of the warping factors for male speakers is expected to be symmetric to that of the female speakers.

The pitch range of all the female speakers in the training data was equally divided into 3 sets: high, medium and low. Similarly, the range of α values derived using the average voice model for these speakers was also divided into 6 equally spaced groups. The warping factors were estimated using the Jacobian normalization; issues of warping factor scaling, use of Jacobian normalization, etc. are not of concern since we are merely interested in dividing speakers into groups.

The distribution of all (40) female speakers present in the training set according to this grouping is shown in Table 4.3. A few significant observations can be made: There is a weak relationship between estimated warping factor and pitch with tendencies of low pitch speakers having lower warping and high pitch speakers to have higher warping. This suggests that most higher pitch voices are associated with females who have shorter vocal tract length.

Natural pitch contours were extracted from the recorded speech of the selected speakers and speech was then resynthesized using the average voice models and the original pitch contours with six different warping factors in the range of 0 to 0.1. Listeners were asked to select the warping factor that synthesized the speech sounding closest to the target speaker. Thus, they selected a single option from the six different versions of the same utterance. Twenty-five listeners were asked to judge the speaker similarity in the speech files synthesized with different warping factors with reference to the natural speech from the speaker. This is repeated for 20 utterances each from a different speaker.

Observations and Discussion: It is interesting to note that combination of a single pitch with different vocal tract lengths can generate a wide variety of voices. A few expert listeners could perceive that the speaker's voice can be almost reproduced from the average voice with the natural pitch and just a single parameter representing the vocal tract length. Even though certain speaker's voices could be almost exactly reproduced, whereas others sounded

Table 4.4 – Correlation between model derived αs (with and without Jacobian normalization) and results of subjective evaluation. The mean, mode and median of the α values are derived from the results of subjective evaluations. Correlation between warping factors from all schemes and pitch is also presented.

	Pitch	**Mean**	**Mode**	**Median**
Pitch	-	-0.1244	-0.1120	-0.0400
Jacobian	-0.4875	0.2238	0.0553	0.2154
No Jacobian	-0.3396	0.4362	0.1976	0.4821

different regardless of the warping factor. This suggests that certain speakers may be better for evaluation. VTLN with a single warping factor cannot replicate the exact speaker identity when the target speaker has speaking styles and accents different from the average voice. If the speaking style and accent of the target speaker is exactly same as the average voice (for example, if both are native British English voices with same speaking style), VTLN can exactly reproduce the speaker characteristics. VTLN will be useful as a speaker transformation even in the accented speech case where, an hierarchical transformation could be trained to accommodate the speaking styles or accent characteristics in a separate transformation (like using a common American English transformation on a UK English model before using VTLN for transforming into an American speaker). For the sake of keeping things simple, different techniques proposed in this work are evaluated using the target speakers who show limited variability in speaking styles compared to the speech from the average voice model.

The results of the subjective evaluations are shown in Figure 4.7. The figure represents the results as a box-plot. Each box represents a speaker in the evaluation set. The lower and upper quartile represent the 25^{th} and 75^{th} percentile and the band in the middle of the box represents the median for the samples. The end points of the whiskers represents the minimum and maximum of the sample values and the outliers are represented by extra points outside the whiskers. It appears that listeners prefer some degree of warping (away from zero warping to give a clearly female voice characteristic). Extreme warping is also not preferred (too “childlike”).

An analysis of the correlation between results from subjective evaluations and α derived from the HMM models is presented in Table 4.4. The α values are derived from the models both with and without Jacobian normalization, which gives a different range for the warping factor distributions (as discussed in Section 4.2.1, the warping factors without Jacobian normalization have a higher range of α values). The table compares the values of mean, mode and median of the warping factors observed in the subjective evaluation. There is no strong correlation between any values. The best correlation is seen between the median of the warping factors from subjective evaluation and those derived from the model without using Jacobian normalization.

Pitch does not show strong correlation to warping factors derived using any scheme. The model derived warping factors have closer correlation to pitch than the warping factors derived

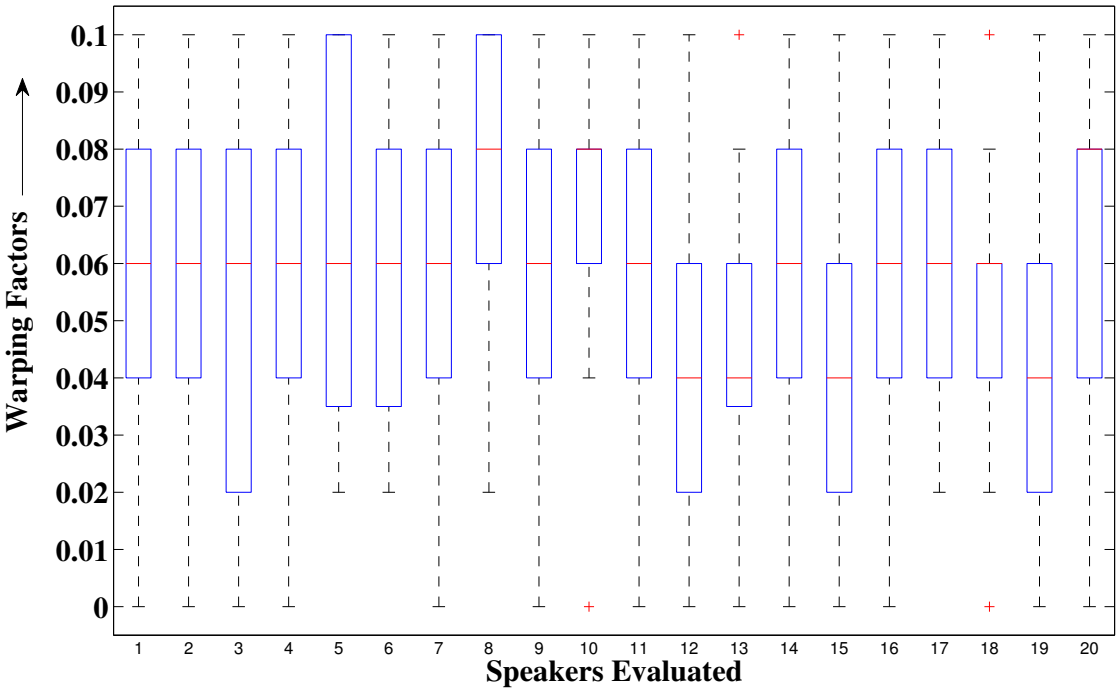


Figure 4.7 – Results from perceptual experiments. The ordinate represents the negative warping factor used to synthesize speech with female characteristics.

from the subjective evaluations.

It is clear from these initial results that perception of VTL is a very difficult task to assess, but it is evident that, on average, the preferred warping factors correspond to a distinctly female voice over one that remains close to the average voice. It is also clear that neither of the two baseline approaches (with and without Jacobian normalization) give impressive results. Thus, investigation of additional techniques as presented earlier has been further motivated. Towards this end, the experiment provides a good prior distribution of the warping factors for VTLN synthesis.

Evaluating VTLN performance

The VTLN transformation aims to perform frequency warping by primarily shifting the formant peaks, and hence, very few speaker characteristics of the target speaker are introduced in the synthesized speech. If a target speaker has very different speaker characteristics when compared to the average voice model, owing to, for instance, accent, dialect, tempo or prosody, it can be difficult to evaluate the characteristics captured by VTLN. Furthermore, it is difficult to guarantee that listeners will evaluate speaker similarity based solely on the specified criterion, further skewing results. The experiments performed in this section, will also help to determine any correlation (if exists) between the different objective and subjective evaluation criteria.

The aim of this research is to evaluate VTLN as a spectral transformation ignoring other factors

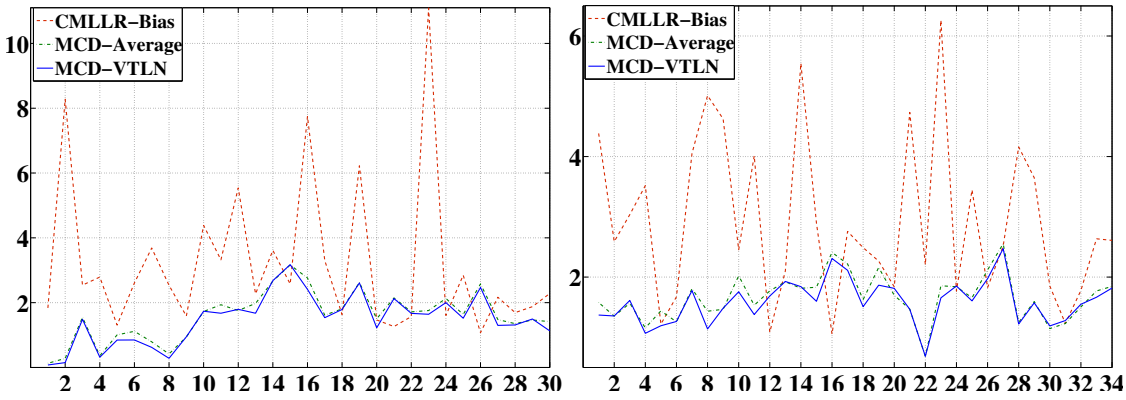
that cannot be captured by VTLN. To this end, experiments were performed to find speakers who can demonstrate characteristics captured by VTLN, and thus can be used to evaluate different VTLN adaptation techniques. Speakers were initially selected using two objective measures and then finally evaluated using subjective scores to find the best target speakers. Experiments were performed on two different English databases, WSJ0 (American English) and WSJCAM0 (British English). Average voice models were built separately using 83 and 92 speakers respectively in the two databases. The aim of this experiment is to find a male and female target speaker from the 30 and 34 selected test speakers of WSJ0 and WSJCAM0 respectively.

It is postulated that adaptation works better on speakers closer to the average voice model [Yamagishi et al., 2010]. The objective scores are chosen so as to find the speakers closer to average voice. The first objective score is based on the bias term in MLLT adaptation techniques, which should ideally represent the displacement of the speaker parameters from the average voice model. CMLLR transforms were generated for the target speakers using the average voice models, and the magnitude of the bias vector was calculated for each speaker. The second objective score is based on the mel-cepstral distortion (MCD) value which is the Euclidean distance between cepstra.

Speech parameters were generated from the average voice using VTLN adaptation for each of the test speakers, and MCD scores were calculated between the synthesized parameters and the parameters extracted from the natural speech of the target speaker. Both MCD and CMLLR-bias values for the speakers from two databases are plotted in Figure 4.8. The figure also shows the MCD scores between average voice and the natural speech of the speaker (denoted as "MCD-Average") which displays the same trend as the score between VTLN synthesized voice and natural speech (denoted as "MCD-VTLN"). It is noted that there is not much correlation between the two objective scores and it is not feasible to select target speakers only based on these scores.

Subjective evaluations were designed in order to find speakers that are rated better with VTLN adaptation. Since it is onerous to perform evaluations on all the test speakers, 10 test speakers were selected from each database. 5 speakers with low MCD scores and 5 with high MCD scores were considered, taking account of gender balance. Listeners were asked to rate the synthesized speech based on speaker similarity on a 5 point scale with 5 being "Sounds like exactly the same speaker" and 1 being "Sounds like an entirely different speaker".

Results plotted as mean opinion scores (MOS) are presented in Figure 4.9. The same speaker indices are used in both subjective and objective plots. The results from these evaluations are summarized in Table 4.5. Speakers are marked as "1" (preferred) and "0" (not preferred) accordingly in the table for each score. It is noted that the subjective results do not have much correlation with any of the objective results. Both scores agree for only a single speaker in both databases. Thus, it can be concluded objective scores may not be adequate to represent absolute VTLN performance. They can however give some insight into relative performance of



(a) WSJ0 Objective (b) WSJCAM0 Objective

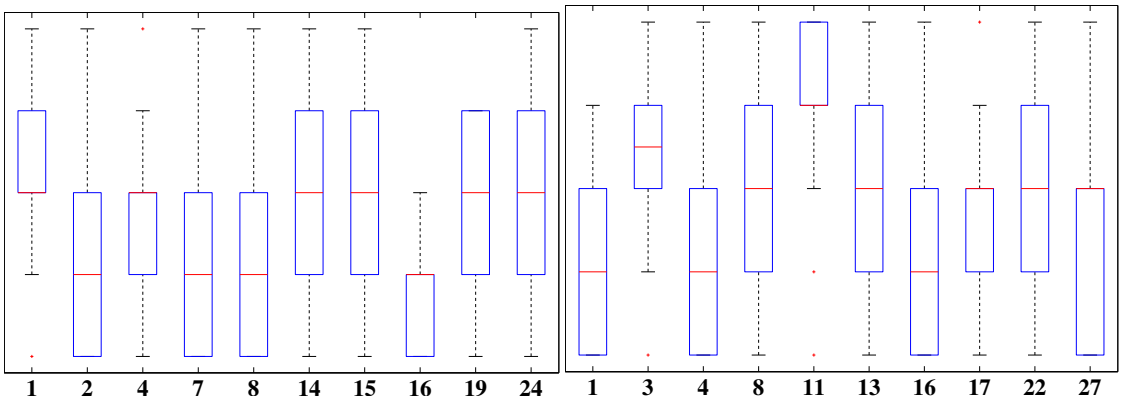
Figure 4.8 – Objective scores for closeness of a speaker to an average voice. Abscissa represents the index of the test speakers and ordinate represents the value for MCD or magnitude of the bias term.

different adaptation techniques especially when the performance is very different. Subjective tests must be performed to efficiently evaluate the VTLN performance in statistical speech synthesis. Also, it is very tricky to find the correct speakers and set-up that can demonstrate the VTLN performance. This motivates the VTLN evaluation in different scenarios as presented in Chapter 7.

In order to perform subjective evaluations with the WSJ0 and WSJCAM0 database and subjectively evaluate different techniques presented in this chapter, a male and a female test speaker with the best subjective score is selected for each database. Conflicts are resolved by a voting scheme between the scores giving least priority to CMLLR-bias and highest priority to subjective results. The next chapter presents the details of the experiments performed to evaluate VTLN techniques and the corresponding results.

4.5 Summary of Contributions

This chapter addressed the divergence of the theory and practice of VTLN implementation, with particular emphasis on the challenges in statistical parametric speech synthesis. Most of the challenges presented in this chapter, are likely to be present in ASR as well, just not prominent enough to have attracted much attention from research community. It follows that the findings in this work should be equally applicable to ASR and help address some questions that have also been raised in ASR (i.e. to use Jacobian or not, to use likelihood scaling or not). It can be argued that the increased sensitivity of TTS to these issues actually helped answer these questions to some extent for ASR as well. Novel solutions were proposed to these challenges. VTLN was placed in a Bayesian setting, which brings in the concept of a prior on the warping factor. The form of the prior, together with acoustic scaling and numerical conditioning were discussed and evaluated. It was seen that the higher order features pose significant problems



(a) WSJ0 Subjective (b) WSJCAM0 Subjective

Figure 4.9 – Subjective scores for WSJ0 and WSJCAM0 databases representing the closeness of a speaker to an average voice. Abscissa represents the index of the test speakers and ordinate represents the DMOS score between 1 and 5.

for the warping factor estimation and hence, an appropriate feature order (of the order 12) representing the spectral envelope should be used to estimate the VTLN warping factor. It can be concluded that the Jacobian determinant was important in VTLN, especially for the high dimensional features used in HMM based speech synthesis. Difficulties normally associated with the Jacobian determinant could be attributed to prior and scaling and use of lower feature dimensionality to estimate warping factors. The experiments showed very little correlation between the subjective and the objective scores for evaluating VTLN and it was not easy to select an appropriate target speaker who can demonstrate its perceptual impact. Most of the work presented in this chapter was published as Saheer et al. [2010b] and Saheer et al. [2012a]

Table 4.5 – Summary of speaker selection experiments for WSJ0 and WSJCAM0. 'M/F' stands for Male/Female speaker. Speakers are classified as having high or low scores. Preferred speakers should ideally have "low" objective scores (CMLLR-Bias & MCD) and "high" subjective scores (MOS). For the sake of improved readability the scores are marked "1" (preferred) and "0" (not preferred) accordingly in the table.

Speaker#	**M/F**	**MOS**	**MCD**	**Bias**
1	**M**	**1**	**1**	**1**
2	F	0	1	0
4	M	1	1	1
7	F	0	1	1
8	F	0	1	1
14	M	1	0	1
15	M	1	0	1
16	F	0	0	0
19	M	1	0	0
24	**F**	**1**	**0**	**1**

Speaker#	**M/F**	**MOS**	**MCD**	**Bias**
1	F	0	1	0
3	F	1	0	0
4	M	0	1	0
8	M	1	1	0
11	**F**	**1**	**1**	**0**
13	M	1	0	1
16	F	0	0	1
17	F	1	0	1
22	**M**	**1**	**1**	**1**
27	M	1	0	1

5 Expectation Maximization based VTLN Implementation

The implementation of VTLN using grid search tends to be computationally expensive due to the extraction of a grid of warped features. Representation of VTLN as a linear transformation in the cepstral domain and then, taking advantage of the model and feature transformation equivalence, the warping factor optimization can be performed in the model domain. This model domain warping factor estimation using gradient descent search based on expectation maximization (EM) can even render VTLN adaptation partially independent of the different feature types.

This chapter presents an EM formulation for the warping factor estimation which improves upon the grid search. This enables more accurate estimation of warping factors and embeds this estimation in the HMM training.

5.1 Related Work

The EM formulation exploits the representation of VTLN as a model transform and does not involve calculation of features with different warping factors. Hence, the warping factors can be estimated very efficiently and precisely. This section presents the background for this refined implementation of VTLN adaptation.

5.1.1 Equivalence of feature & model transformations

It was suggested by Pitz and Ney [2005] that VTLN can be represented as a linear cepstral transformation and can be considered a special case of maximum likelihood linear regression (MLLR) adaptation technique by Leggetter and Woodland [1995]. The transformation matrices were analytically derived for three typical warping functions. These invertible warping functions were piece-wise linear, quadratic and bilinear warping functions. It was shown that the matrices were diagonally dominant and thus, can be approximated with quindiagonal matrices or even with a tridiagonal matrix for MFCCs which already include mel-warping. Further, it was shown that there is a strong interdependence between VTLN and MLLR if

Gaussian emission probabilities are applied.

Assuming the state output probability distribution for an HMM to be Gaussian:

$$p(\boldsymbol{x}_\alpha \mid \boldsymbol{\Theta}) = |\boldsymbol{A}_\alpha| N(\boldsymbol{A}_\alpha \boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{5.1}$$

Assuming a single mixture component and the feature transform representation will yield:

$$p(\boldsymbol{x}_\alpha \mid \boldsymbol{\Theta}) = \frac{|\boldsymbol{A}_\alpha|}{\sqrt{(2\pi)^N |\Sigma|}} \exp -\frac{1}{2} (\boldsymbol{A}_\alpha \boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{A}_\alpha \boldsymbol{x} - \boldsymbol{\mu}) \tag{5.2}$$

Multiplying the power of the exponent term with $\boldsymbol{A}_\alpha \boldsymbol{A}_\alpha^{-1}$

$$p(\boldsymbol{x}_\alpha \mid \boldsymbol{\Theta}) = \frac{|\boldsymbol{A}_\alpha|}{\sqrt{(2\pi)^N |\Sigma|}} \exp -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{A}_\alpha^{-1} \boldsymbol{\mu})^T \boldsymbol{A}_\alpha^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A}_\alpha (\boldsymbol{x} - \boldsymbol{A}_\alpha^{-1} \boldsymbol{\mu}) \tag{5.3}$$

Rearranging terms as a Gaussian equation again:

$$p(\boldsymbol{x}_\alpha \mid \boldsymbol{\Theta}) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{A}_\alpha^{-1} \Sigma (\boldsymbol{A}_\alpha^{-1})^T|}} \exp -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{A}_\alpha^{-1} \boldsymbol{\mu})^T (\boldsymbol{A}_\alpha^{-1} \boldsymbol{\Sigma} (\boldsymbol{A}_\alpha^{-1})^T)^{-1} (\boldsymbol{x} - \boldsymbol{A}_\alpha^{-1} \boldsymbol{\mu}) \tag{5.4}$$

This implies that

$$|\boldsymbol{A}_\alpha| N(\boldsymbol{A}_\alpha \boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv N(\boldsymbol{x} \mid \boldsymbol{A}_\alpha^{-1} \boldsymbol{\mu}, \boldsymbol{A}_\alpha^{-1} \boldsymbol{\Sigma} (\boldsymbol{A}_\alpha^{-1})^T) \tag{5.5}$$

Thus, a linear transformation of the observation vector $\boldsymbol{x}$ is equivalent to a linear transformation of the mean vector μ and appropriate transformation of the covariance matrix, Σ. This transformation is a special case of MLLR, known as constrained MLLR (CMLLR) by Gales and Woodland [1996] where the phrase constrained attributes to the use of same matrix for transformation of mean and variance. Hence, it was shown that CMLLR and VTLN are equivalent in formulation and it was shown empirically by Uebel and Woodland [1999] that the performance improvements with CMLLR need not necessarily be additive to VTLN. It was further shown that the Jacobian term to be considered when directly comparing probability values for differently normalized distributions can be safely approximated as the log determinant of the transformation matrix. This was mainly to keep in check the errors caused by improper normalization of the probability distribution and would not cause any failure for ASR feature normalization. It was mentioned that the reason why VTLN in contrast to CMLLR does not fail without proper normalization is in the limitation of the warping factors and specification of the warping function. VTLN is a constrained form of CMLLR with limited range of feasible warping factors and thus, matrix is closer to an identity matrix.

The general form of the CMLLR transform is:

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}' \boldsymbol{\mu} - \boldsymbol{b}'$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{A}'\boldsymbol{\Sigma}\boldsymbol{A}'^{T} \tag{5.6}$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the transformed mean ($\boldsymbol{\mu}$) and variance ($\boldsymbol{\Sigma}$) of a Gaussian distribution. $\boldsymbol{A}'$ is the transformation matrix which is equivalent to the $\boldsymbol{A}^{-1}$ matrix in Equation 5.5. It was shown by Gales [1998] that representation of CMLLR as a feature transformation can help in the estimation of transformation matrix using an iterative EM solution. The EM auxiliary function could be optimized using the transformed sufficient statistics of the observed feature vectors which alleviates the problems associated with full covariance matrices (generated by multiplying the diagonal covariance matrix with the transformation matrix) typically used in HMM based speech recognition. Similar strategies could be used in the efficient implementation of VTLN.

5.1.2 Earlier Implementations

Though it is an inexact method for performing VTLN, if the warped spectrum could be estimated from the unwarped cepstral features, the full feature computation for each point in a grid search could be bypassed. The development of cepstrum domain VTLN [Pitz, 2005, Umesh et al., 2005, Panchapagesan and Alwan, 2009] helped alleviate this issue further by demonstrating the equivalence of spectral warping and linear cepstrum transformations [Pitz and Ney, 2005]. The knowledge that cepstral transformations can be shown to be equivalent to transformations in the spectrum [Pitz and Ney, 2005] paved the way to the implementation of VTLN in the cepstral domain [Pitz, 2005, Umesh et al., 2005, Panchapagesan and Alwan, 2009]. VTLN has also been implemented in other feature domains such as perceptual linear prediction (PLP) features [McDonough, 2000] for ASR. A break-through in this field was the use of the expectation-maximization (EM) [Dempster et al., 1977] formulation, which improved efficiency through the use of a gradient descent search (on a grid of warping factor values). Most of these implementations still used a grid of pre-computed transformation matrices or a simple (linear) warping function to estimate a tractable gradient for the EM auxiliary function.

EM is well-known and widely used for optimization problems which can be formulated as instances of maximum likelihood parameter estimation. It has been shown that EM can be used to estimate VTLN warping factors for ASR [McDonough, 2000, Panchapagesan and Alwan, 2009, Akhil et al., 2008]. Warping parameters are estimated by maximizing the EM auxiliary function over the adaptation data. The objective function obtained is similar to the one used in MLLR or CMLLR [Gales, 1998]. The same sufficient statistics as used in CMLLR can be used for optimizing the VTLN auxiliary function.

A set of precomputed α matrices can be multiplied with the sufficient statistics to estimate the optimal warping factors [Panchapagesan and Alwan, 2009]. This approach reduces to a grid search rather than gradient descent estimation. Panchapagesan and Alwan [2009] proposed a compact representation of linear transformation for frequency warping that can be calculated

for any frequency warping function and showed the equivalence between different linear transformations proposed earlier. The EM auxiliary function based on maximum likelihood criterion was optimized using the same sufficient statistics as CMLLR for a grid of possible warping factors. A similar approach was presented by Akhil et al. [2008] that used EM to compare the likelihoods of the warped features computed by appropriately modifying the sufficient statistics using the warp matrices. This work used a linear transform based VTLN warping applied directly on the cepstral features [Rath et al., 2009]. A soft alignment based on the forward backward algorithm in HMM is used to generate the expected values of the hidden variables (the state and mixture component posterior probabilities). These values can then be used to calculate conditional likelihoods of the model for different warping factors and choose the best warping factor (the M-step of the EM algorithm). A grid of pre-computed warping matrices are used in this case as well.

Regression class formulation for maximum likelihood linear transformations (MLLTs) enables estimation of different transforms for different Gaussian mixture components in the model where the number of transforms generated is dictated by the quantity of available adaptation data. Applying this to EM-based VTLN permits application of different warping factors for different classes. This approach was used by Panchapagesan and Alwan [2009], Rath and Umesh [2009]. Rath and Umesh [2009] built regression classes using both data driven approach and using phonetic knowledge. Again, a set of pre-computed warping matrices were used for estimating the best warping factor in the EM framework. The use of regression class trees ensured that the warping factors were estimated for all classes even when the data is insufficient in which case, the parent class was used for warping factor estimation. The study showed that the different phoneme classes demonstrated differences in the warping factor especially that represented vowels and consonants. The work found that the performance improved when the silence class was omitted from the warping.

Similar performance improvements were demonstrated with multiple VTLN transformations using both the data-driven and phonetic knowledge based regression classes. It was also shown by Miguel et al. [2005] that multiple spectral transformations can improve performance. The work proposed an augmented state space acoustic decoder named "MATE" with a locally constrained search for spectral warping that is augmented into the optimal state sequence search in the standard Viterbi algorithm. It was argued that all phonetic events do not exhibit similar spectral variation as a result of physiological differences on vocal tract shape. Hence, it is ideal to use different warping factors for the different phonetic units. The Viterbi algorithm was applied on a 3-dimensional trellis, where the third dimension represents the ensemble of possible values for the warping factors. This augmented state space decoder also used transition probabilities that constrained the frequency warping transformations applied to the adjacent frames to be taken from the adjacent indices of the ensemble (grid). This resulted in a decoder that reduces the impact of local spectral and temporal variabilities in ASR performance. This technique was also incorporated into the HMM framework to generate models with speaker mismatch reduction [Miguel et al., 2006, 2008]. This same system was used to show improvements in the pathological speech recognition by Saz et al. [2006].

Higher order terms in the VTLN matrix can be ignored to give a closed form solution for the EM formulation as shown by Emori and Shinoda [2001], Hirohata et al. [2003]. Hirohata et al. [2003] used only the first four cepstral coefficients in the EM optimization to restrict the gradient of the auxiliary function to be linear. Optimization using simple terms in the matrix or using few lower order cepstral coefficients does not guarantee that the estimated α will maximize the likelihood over the entire feature vector. It can be note that the dimensionality of order four (used by Hirohata et al. [2003]) is not sufficient to represent the spectral envelope which needs to be warped using VTLN (see also Figure 4.5) .

One of the pioneers in implementing an efficient EM formulation for VTLN was McDonough [2000]. All-pass transform based VTLN was formulated using a maximum likelihood criterion based auxiliary function. Formulations for partial gradients and partial Hessian were also derived. The simple bilinear transform based warping with a single parameter was optimized using Brent's search. More complicated all-pass transforms with multiple parameters were optimized using Newton's method. Inspired by the work of McDonough [2000], this research presents an algorithm based on Brent's search for finding the optimal value of the warping factor from this auxiliary function.

5.2 Expectation Maximization implementation

Any general VTLN implementation faces two main challenges, (1) efficient calculation of VTL-warped features for a given warping factor and (2) optimization of the warping factor with low time and space overhead. In addition, application of VTLN for TTS requires addressing of issues specific to this domain.

The first of these issues can be addressed by implementing VTLN as a cepstral transformation. It is argued persuasively by Pitz and Ney [2005] that VTLN amounts to a linear transform in the cepstral domain. In fact, this is also evident from the mel-generalized approach to feature extraction [Tokuda et al., 1994b], with the use of bilinear warping function being of particular interest due to its presence in the mel-frequency warping function of MGCEP features. It was demonstrated in Chapter 3 (Section 3.2) that all-pass transform warping for MGCEP can in fact be represented as a linear cepstral transformation. Hence, VTLN can also be implemented as an equivalent model transform, similar to CMLLR.

The second issue is addressed through formulation of warping factor estimation in the framework of expectation maximization. Representation of VTLN as a model transformation enables the use of EM for finding the optimal warping factors [McDonough, 2000, Akhil et al., 2008, Panchapagesan and Alwan, 2009]. This provides advantages in terms of more precise estimation of warping factor, α, and improved efficiency in time and space. EM can be embedded into the HMM training utilizing the same sufficient statistics as CMLLR. This also opens up the possibility of estimating multiple warping factors for different phone classes. Since the ML optimization does not provide a closed form solution to the EM auxiliary function, Brent's search is used to estimate the optimal warping factors.

5.2.1 EM auxiliary function

Similar to CMLLR adaptation, the feature transform can be analogously represented as a model transform [Gales, 1998]. The maximum likelihood optimization for a Gaussian distribution in the feature domain is:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} |\boldsymbol{A}_\alpha| N(\boldsymbol{A}_\alpha \boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{5.7}$$

The same equation can be represented as a model transform:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} N(\boldsymbol{x} \mid \boldsymbol{A}_\alpha^{-1}\boldsymbol{\mu}, \boldsymbol{A}_\alpha^{-1}\boldsymbol{\Sigma}(\boldsymbol{A}_\alpha^{-1})^T) \tag{5.8}$$

where $\boldsymbol{x}$ represents the cepstral feature vector, μ and Σ correspond to the mean and variance of the Gaussian model. The Jacobian normalization can be calculated as the determinant of the matrix $\boldsymbol{A}_\alpha$ representing the linear transformation of the cepstral features.

The maximum a posteriori (MAP) criterion for warping factor estimation similar to the ML criterion introduced in Equation 5.7 can be represented as follows:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} |\boldsymbol{A}_\alpha| p(\boldsymbol{x}_\alpha \mid \boldsymbol{\Theta}, w) p(\alpha \mid \boldsymbol{\Theta}) \tag{5.9}$$

where, $p(\alpha \mid \Theta)$ is the prior probability of α for a given model, $\boldsymbol{\Theta}$. When using the likelihood comparison to search for the best warping factor, Jacobian normalization has to be taken into consideration [Pitz, 2005, Sankar and Lee, 1996]. The EM formulation of warping factor estimation results in the following auxiliary function. Taking the log of the function and considering the Gaussian assumption for the model and marginalizing out the state sequences (hidden variable).

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \Big\{ \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_{m,f} \Big[\log(N(\boldsymbol{A}_\alpha x_f \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)) + \log|\boldsymbol{A}_\alpha| \Big] + \log p(\alpha \mid \boldsymbol{\Theta}) \Big\} \tag{5.10}$$

where, $\boldsymbol{A}_\alpha$ is the transformation matrix for input feature vector $\boldsymbol{x}$, M is the total number of mixtures, F is the total number of frames, γ_m is the posterior probability of mixture m, and $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are the parameters of the Gaussian mixture component, m.

Expanding the terms in the above equation, estimation of a warping factor using this criterion can be shown to be equivalent to maximizing the following auxiliary function [McDonough, 2000].

$$Q(\alpha) = \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_{m,f} \log\Big(\frac{1}{\sqrt{(2\pi)^N |\Sigma_m|}} \exp -\frac{1}{2} (\boldsymbol{A}_\alpha \boldsymbol{x}_f - \boldsymbol{\mu}_m)^T {\boldsymbol{\Sigma}_m}^{-1} (\boldsymbol{A}_\alpha \boldsymbol{x}_f - \boldsymbol{\mu}_m) \Big)$$
$$+ \beta \log|\boldsymbol{A}_\alpha| + \log p(\alpha \mid \boldsymbol{\Theta}) \tag{5.11}$$

where, N is the dimensionality of the features and,

$$\beta = \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_{m,f}$$

In the case of a single mixture, β could reduce to F, the total number of frames. Further expanding and setting the terms independent of warping factor α to K,

$$Q(\alpha) = \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_{m,f} \left[-\frac{1}{2} (\boldsymbol{A}_\alpha \boldsymbol{x}_f - \boldsymbol{\mu}_m)^T {\boldsymbol{\Sigma}_m}^{-1} (\boldsymbol{A}_\alpha \boldsymbol{x}_f - \boldsymbol{\mu}_m) \right] + \beta \log |\boldsymbol{A}_\alpha| + \log p(\alpha \mid \boldsymbol{\Theta}) + K \quad (5.12)$$

Assuming a diagonal covariance for the auxiliary function results in maximization of the following function.

$$Q(\alpha) = -\frac{1}{2} \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_{m,f} \sum_{i=1}^{N} \frac{(\boldsymbol{A}_{\alpha_i} \boldsymbol{x}_{f_i} - \boldsymbol{\mu}_{m_i})^2}{\boldsymbol{\sigma}_{m_i}^2} + \beta \log |\boldsymbol{A}_\alpha| + \log p(\alpha \mid \boldsymbol{\Theta}) + K \quad (5.13)$$

The auxiliary function represented by EM (Equation 5.13) can use statistics similar to the CMLLR estimation as derived by Gales [1998]. It results in the following auxiliary function.

$$Q(\alpha) = -\frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{w}_i \boldsymbol{G}_i \boldsymbol{w}_i^T - 2 \boldsymbol{w}_i \boldsymbol{k}_i^T) + \beta \log |\boldsymbol{A}_\alpha| + \log p(\alpha \mid \boldsymbol{\Theta}) + K \quad (5.14)$$

where,

$$\boldsymbol{G}_i = \sum_{m=1}^{M} \frac{1}{\boldsymbol{\sigma}_{m_i}^2} \sum_{f=1}^{F} \gamma_{m,f} \boldsymbol{x}_f \boldsymbol{x}_f^T \quad (5.15)$$

$$\boldsymbol{k}_i = \sum_{m=1}^{M} \frac{1}{\boldsymbol{\sigma}_{m_i}^2} \boldsymbol{\mu}_{m_i} \sum_{f=1}^{F} \gamma_{m,f} \boldsymbol{x}_f^T \quad (5.16)$$

and $\boldsymbol{w}_i$ represents the i^{th} row of the transformation matrix $\boldsymbol{A}_\alpha$.

The objective function demands differentiation of the determinant of the transformation matrix for a closed form solution. Furthermore, all the terms in the VTLN transformation matrix are interrelated (functions of α) and a co-factor approach similar to the EM optimization for CMLLR is not feasible. A Brent's search technique as explained in section 5.2.2 is used to find the best warping factor using this auxiliary function. For VTLN, the bracket of the search is bounded by -0.1 and 0.1.

5.2.2 Brent's search

In this work, Brent's search [Press et al., 1992] is used to find an optimal warping factor for the EM auxiliary function of Equation 5.13. Brent's search is a method for general one-dimensional root finding. This method combines root bracketing, bisection and inverse quadratic interpolation to converge from the neighbourhood of zero crossing. The inverse quadratic interpolation uses three prior points to fit an inverse quadratic function. If x is represented as a quadratic function of y, setting $y = 0$ gives the next estimate of the root x. The search guarantees that the root lies within the specified bracket. For VTLN, the bracket of the search is bounded by -0.1 and 0.1. Bounds iteratively collapse until the desired minimum is reached. A bracketing triplet ((a, b, c) such that function value f(b) is less than or equal to both f(a) and f(b)) is used for finding the root of the desired function. When the function has a proper minimum (i.e. function represents a parabola), the triplets can be used for fitting the quadratic equation using the inverse parabolic interpolation which is the best method of optimization in this case. With a most uncooperative function, the golden section search is used to find the minimum, where the next set of bounds are estimated using the golden section ratio (fraction equal to 0.38197) of the bigger segment, converging to a proper self-replicating ratio. If the bounds are not collapsing rapidly, the algorithm uses a bisection step. Brent's method combines the sureness of bisection with the speed of the higher order method when appropriate. The main advantage of using Brent's search is that no derivative needs to be computed.

5.2.3 Full MAP Estimation

The EM formulation presented here is a novel technique based on the MAP optimization criterion shown in Equation 5.9 instead of the ML optimization shown in Equation 2.2. There is a prior term $p(\alpha \mid \Theta)$ which determines the probable distribution of the warping factors depending on the model set. A general approach is to ignore this term or set it flat resulting in an assumption of uniform distribution. While training, this term may not have any effect since there is a lot of data available for training and the prior is insignificant compared to the amount of data (likelihood scores). During testing, when the warping factor is to be estimated from a single utterance, this term could be more valuable. In the case of synthesis (as presented in chapter 4), the challenge is to estimate the desired distribution of the warping factors which can be perceived as acceptable.

5.2.4 Class-based multiple transforms

VTLN is generally implemented using a single warping factor for an entire utterance or most often all the utterances of a single speaker representing a global spectral warping. However, Rath and Umesh [2009] showed that not all phonemes exhibit the same spectral variation due to physiological differences. Some phoneme classes like the vowels or semi-vowels can be assumed to have a better representation of formants in the spectral envelope when compared to

other classes like fricatives or unvoiced consonants. Consonants, and especially unvoiced consonants involve constrictions in the vocal tract that means that the "length" of the vocal tract is being deformed during speech production. This is also a major reason why the warping factors can change across different classes of sounds. The silence class will not have any formant structure at all and need not be warped. It should be ideal to use different warping factors for different phone classes. Multiple warping factors have yielded improvements in recognition performance. Data can be divided into acoustic classes using data-driven approaches or using phonetic knowledge as shown by Rath and Umesh [2009]. Phone dependent warping can be implemented after obtaining phone labels from a first pass recognition [Molau et al., 2000]. Frame specific warping factors can also be estimated by expanding the HMM state space with some constraints [Miguel et al., 2005].

In speech synthesis, phone classes can also be synthesized with different warping factors for a single speaker. Using similar statistics as CMLLR helps the implementation of EM-VTLN in standard speech processing toolkits like HTK. Multiple MLLTs are usually applied using a regression class tree. Such regression classes can also be employed in multi-class VTLN. The regression class tree structure is derived from the decision tree clustering as in HTS [Yamagishi et al., 2009b]. The CMLLR transforms usually use regression tree based classes to estimate transforms for groups of units. Implementing VTLN like a CMLLR transform estimation helps in using the regression class based transform estimation for VTLN. Each regression class can have different warping factors. This can result in different warping for different classes resulting in appropriate warping for each sound according to factors like place of articulation. This can result in less warping for unvoiced units and more warping for voiced components. This research investigates multi-class EM-VTLN estimation in the context of statistical synthesis. Unlike ASR, there are issues with VTLN estimation for statistical speech synthesis which were investigated in the chapter 4.

5.3 Experiments & Results

As explained in earlier chapters, VTLN implementation poses a number of challenges especially in the framework of statistical parametric speech synthesis. This section presents experiments carried out to evaluate the performance of different VTLN approaches in statistical parametric speech synthesis. More specifically, the baseline all-pass transform-based VTLN using EM learning criterion, as presented in Section 5.2, is compared against several variants based on the methods presented in Chapter 4. The goal of this work is to find the most effective VTLN approach for statistical parametric speech synthesis and, by way of this, explain some of the differences observed in comparing past literature on VTLN to this one. In comparing different VTLN estimation techniques for synthesis it is not aimed to surpass the performance of more powerful adaptation approaches such as CSMAPLR or CMLLR. Instead, the longer-term goal is to combine VTLN with such adaptation approaches, as will be discussed further in the later chapters.

Table 5.1 – Techniques to be evaluated for VTLN warping factor estimation

Name	Technique
T1	Jacobian Normalization
T2	No Jacobian Normalization
T3	Jacobian Normalization with scaled prior
T4	Jacobian Normalization with scaled likelihood
T5	Combination of T3 and T4
T6	Using feature blocksize 13 & Jacobian Normalization
T7	Using feature blocksize 13 & No Jacobian Normalization

All-pass transform based VTLN using ML optimization was implemented using the EM formulation and embedded into the HMM training. Finally, speech was generated using transformed model parameters. Evaluations were performed using two databases - WSJ0 (American-English) and WSJCAM0 (British-English) databases. WSJ0 has 83 training speakers while WSJCAM0 has training 92 speakers. Selecting target speakers for evaluating VTLN is not an easy task. Target speakers selected using experiments presented in the previous chapter (Chapter 4) were used to evaluate different techniques for estimating warping factors.

5.3.1 Techniques evaluated

As explained in Chapter 4, there are various techniques to overcome the challenges in estimating warping factors for TTS. The approaches discussed were primarily introduced to address higher TTS feature dimensionality and included:

– **Omission of Jacobian normalization** as has previously been done in ASR studies;
– **Prior distribution of warping factors** to provide a better estimate in test conditions when there is little adaptation data;
– **Likelihood scaling** applied to acoustic scores (while omitting scaling for the Jacobian); and
– **Lower order features** for the estimation of warping factor, thereby ignoring high order cepstra in EM auxiliary function;

All the techniques evaluated in this section are summarized in Table 5.1. The same nomenclature as in the table is used when labeling the results. The scale factors for prior (value of 10) and likelihood (value of 3) were estimated empirically in cases (b) and (c) above (for T3, T4 and T5 in Table 5.1). Evaluations were also performed for both single and multiple (regression class-based) transform VTLN. It would have been ideal to estimate different prior distributions for each class in a multi-class VTLN system. Since this is out of the current scope of this research, same prior distributions are used for different regression classes in the multi-class VTLN system. This is one of the possible future directions for this work.

Table 5.2 – MCD (in dB) for VTLN synthesis for WSJ0 with 10 test speakers.AV represents average voice.

Blocksize	**Jacobian**	**Scaling**	**Global**	**Multiple**
AV	-	-	7.616	-
39	Yes	None	7.518	7.517
39	No	None	7.469	7.464
39	Yes	Prior	7.439	7.500
39	Yes	Likelihood	7.464	7.469
39	Yes	Combination	7.441	7.465
13	Yes	None	7.456	7.433
13	No	None	7.527	7.507

Table 5.3 – MCD (in dB) for VTLN synthesis for WSJCAM0 with 10 test speakers.AV represents average voice.

Blocksize	**Jacobian**	**Scaling**	**Global**	**Multiple**
AV	-	-	6.107	-
39	Yes	None	5.994	5.982
39	No	None	6.000	5.980
39	Yes	Prior	5.973	5.964
39	Yes	Likelihood	6.010	6.018
39	Yes	Combination	5.988	5.946
13	Yes	None	5.986	5.973
13	No	None	5.977	5.975

5.3.2 Experimental Setup

The HMM speech synthesis system (HTS), more specifically the system scripts for the HTS-2007 submission to the Blizzard Challenge [Yamagishi et al., 2009b] provided the basis for training and generating the statistical parameters for speech synthesis. HTS models spectrum, pitch (log f0), band-energy and duration in the unified framework of hidden semi-Markov models (HSMMs). Features extracted were 39th-order mel-cepstra [1] derived from STRAIGHT spectrum, log F0, five-band aperiodicity, and their delta and delta-delta coefficients, extracted from 16kHz recordings with a window shift of 5ms. The STRAIGHT vocoder was used to synthesize speech from the parameters generated using HTS.

Two English average voice synthesis model sets were trained on the WSJ0 and WSJCAM0 corpora using an HMM five-state left-to-right topology. Models were trained using CMLLR-based speaker adaptive training (CMLLR-SAT), while all feature extraction was performed with $\alpha = \alpha_M = 0.42$, which was applied to the untruncated cepstrum. This approach was used in order to facilitate the evaluation of different VTLN estimation techniques using a common canonical model. For synthesis, the VTLN warping factor, α_V, was estimated as a

1. SPTK's `mcep` function was used, which is equivalent to mel-generalized ceptrum with $\gamma = 0$ while it does not apply the UELS optimization criterion.

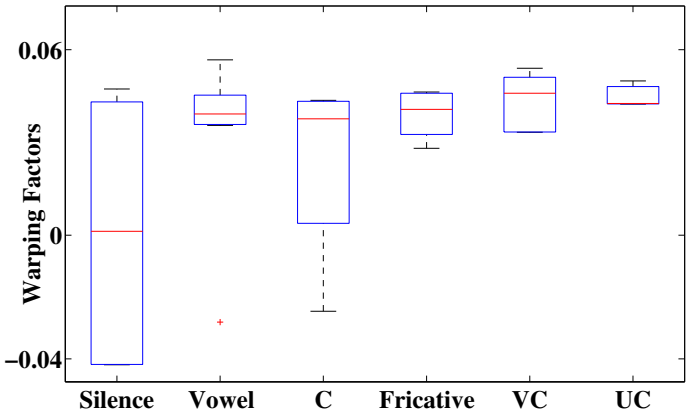


Figure 5.1 – Distribution of α for different phoneme classes (C = all consonants, VC = voiced consonants, UC = unvoiced consonants) for a specific male speaker. The global VTLN warping factor in this case is "0.0467". Note that C $\neq$ VC $\cup$ UC; see the text.

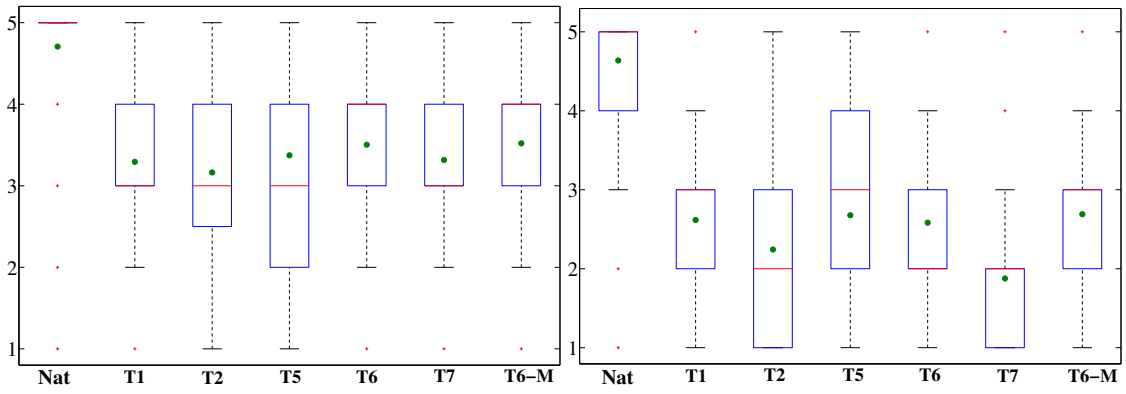
model transform using EM approach as described in this chapter. However, for synthesis, the negative of α_V was used as a model transform (implemented by multiplying $A_{-\alpha}$ matrix with the 40 dimensional model parameters and synthesizing features with the transformed models).

For objective scores, different VTLN warping factor estimation techniques were evaluated for ten (5 male and 5 female) target speakers from each database. Only a single sentence from each target speaker was used as the adaptation data. The objective measures are based on mel-cepstral distortion (MCD). MCD is the Euclidean distance of synthesized cepstra from that of the values derived from the natural speech.

The subjective performance was evaluated based on naturalness and speaker similarity using MOS and DMOS scoring respectively. The synthesized utterances were rated on a 5-point scale, 5 being “completely natural” or “sounds like the exact target speaker” and 1 being “completely unnatural” or “sounds like a totally different speaker”. For subjective testing, one male and one female test speaker from each database were selected based on experiments presented in 4.4.1 (previous chapter). The subjective evaluations were performed using Amazon Mechanical Turk [2] online evaluation setup. The listeners were paid for their service.

5.3.3 Results and Discussion

This section presents the results of the evaluations and discusses some of the conclusions that can be drawn from these results.



(a) WSJ0 Database (b) WSJCAM0 Database

Figure 5.2 – Subjective Scores for Naturalness. Nat represents Vocoded speech and all VTLN systems use only a single parameter except for T6-M system which is the multiple parameter version of the T6 system.

Table 5.4 – Wilcoxon signed rank test for significance of 1% for Naturalness for WSJ0 and WSJCAM0.

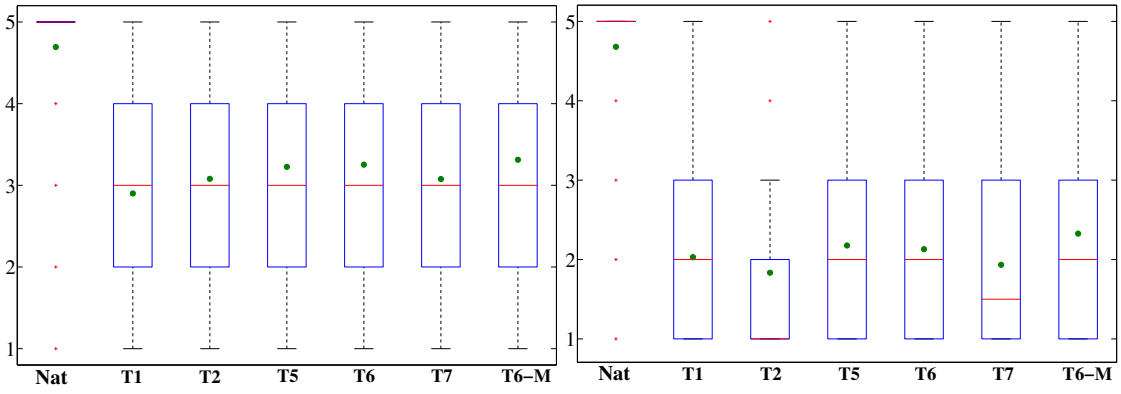
	Nat	T1	T2	T5	T6	T7	T6-M
Nat	-	1	1	1	1	1	1
T1	1	-	0	0	1	0	1
T2	1	0	-	1	1	0	1
T5	1	0	1	-	0	0	0
T6	1	1	1	0	-	1	0
T7	1	0	0	0	1	-	1
T6-M	1	1	1	0	0	1	-

	Nat	T1	T2	T5	T6	T7	T6-M
Nat	-	1	1	1	1	1	1
T1	1	-	1	0	0	1	0
T2	1	1	-	1	1	1	1
T5	1	0	1	-	0	1	0
T6	1	0	1	0	-	1	0
T7	1	1	1	1	1	-	1
T6-M	1	0	1	0	0	1	-

Analysis of Warping Factors

An analysis of warping factors obtained by the proposed approach for multi-class VTLN is presented in this section. The distribution of α for different phoneme classes for a male speaker is shown in the Figure 5.1. The values were derived using Jacobian normalization with scaled likelihood and prior. Using this method results in warping factors that are slightly biased towards the prior for all classes which explains the high warping factors for some consonant classes. A clearer difference would be observed between classes in the case where no prior is used. It is observed that silence has very noisy warping factors and ideally should be ignored in adaptation. Multi-class VTLN can facilitate this task by ignoring the classes representing silence. Consonants, in general, display warping factors tending to lower than average values. Whilst the consonants class represents all consonants, the voiced and unvoiced categories of consonants shown in the figure are the small subsets that had *only* those labels, rather than combinations of those and other labels such as pulmonic or plosive. These two pure classes show somewhat opposite trends to one another.

2. https://www.mturk.com



(a) WSJ0 Database (b) WSJCAM0 Database

Figure 5.3 – Subjective Scores for Speaker Similarity. Nat represents Vocoded speech and all VTLN systems use only a single parameter except for T6-M system which is the multiple parameter version of the T6 system.

Table 5.5 – Wilcoxon signed rank test for significance of 1% for Speaker similarity for WSJ0 and WSJCAM0.

	Nat	T1	T2	T5	T6	T7	T6-M
Nat	-	1	1	1	1	1	1
T1	1	-	0	1	1	0	1
T2	1	0	-	0	0	0	1
T5	1	1	0	-	0	0	0
T6	1	1	0	0	-	0	0
T7	1	0	0	0	0	-	1
T6-M	1	1	1	0	0	1	-

	Nat	T1	T2	T5	T6	T7	T6-M
Nat	-	1	1	1	1	1	1
T1	1	-	1	0	0	0	1
T2	1	1	-	1	1	0	1
T5	1	0	1	-	0	1	0
T6	1	0	1	0	-	1	0
T7	1	0	0	1	1	-	1
T6-M	1	1	1	0	0	1	-

Objective Results

The objective evaluations based on MCD are performed for all the 10 speakers selected for the experiments in section 4.4.1. All the techniques proposed in this chapter are evaluated objectively. The abbreviations for the systems are presented in Table 5.1 and MCD results in Table 5.2 and Table 5.3. It was observed that a few speakers/systems give extreme warping factors which result in huge MCD scores which was fixed by using $A_{-\alpha}$ instead of A_{α}^{-1} during synthesis. Systems that do not use Jacobian normalization have relatively higher MCD scores and hence, are not preferred. The systems that use lower order features or use prior and likelihood scaling also perform better. Even after generating a desired warping factor distribution, the system T4 (Jacobian normalisation with scaled likelihood) for the WSJCAM0 system shown in Table 5.3 is an anomaly to this observation. This is because objective scores are not always consistent with the subjective observations and the scale factor is empirically tuned with a different set of data.

Subjective Results

In order to simplify the listening tests, not all techniques are subjectively evaluated. Techniques to be evaluated subjectively are selected based on the objective results. Global transform versions of all the VTLN systems presented in Table 5.1 were evaluated. Apart from these, the T6-M system was also evaluated. The T6-M system represents the multiple transform case for the T6 system. From the observations based on earlier research [Saheer et al., 2010a] and the objective results, T3 and T4 systems are not very different from their combination T5 system. Hence, these systems are omitted from subjective evaluation. 75 listeners participated in the evaluation and listened to 124 sentences in total. The results for naturalness and speaker similarity are plotted in Figure 5.3. Table 5.4 and Table 5.5 shows if the systems are significantly different based on the Wilcoxon signed rank test (with significance of 1%) [3]. A few systems are shown to be significantly different from each other. Listeners do not prefer systems that did not use Jacobian normalization. The systems that use lower order features to estimate the VTLN warping factors are preferred in general (both while using single and multiple VTLN parameters).

Both objective and subjective results for systems not using Jacobian normalization (systems T2 and T7) are worse when compared to other systems. This can be attributed to the fact that not using Jacobian normalization not only gives higher values for warping factors, but also sometimes gives extreme values (as shown in Figure 4.1d). This extreme warping is not preferable for the test speakers when evaluating speaker similarity or naturalness and also results in the huge MCD values for these speakers unless the inverse of the α matrix is calculated carefully. This same phenomenon was observed when using the likelihood scaling (system T4) for some test speakers. It can be concluded from these results that Jacobian normalization is an important factor that should not be avoided for estimating VTLN warping factors, especially when using a large number of test speakers, who might give some extreme results when not using Jacobian.

However, at the same time, it can be observed that using Jacobian normalization directly with higher order features (system T1) is not very preferable especially for speaker similarity or MCD scores. Using Jacobian normalization with higher order features over compensates the variations in the warping factors and reduces the spread of warping factors. This in turn results in less discrimination among different test speakers and further reduces the speaker similarity. These observations motivate the use a technique that can overcome the problem of feature dimensionality such as using the lower order features for estimating the warping factor (systems T6 or T7), which is a novel technique presented in this work.

From the above experiments, it can be concluded that the best method to estimate VTLN warping factors for statistical parametric speech synthesis is using lower order features with

3. The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two samples to assess whether the median difference between pairs of observations is zero (i.e. it is a paired difference test). It can be used as an alternative to the paired Student's t-test when the population cannot be assumed to be normally distributed.

Jacobian normalization (system T6). This technique gives stable objective results for all speakers and comparatively better subjective evaluation performance. Though there were minor improvements in the objective scores, the multiple transform case (using regression class trees) for this system could not display any statistically significant improvement in the performance over the global transform based VTLN adaptation. The MAP based VTLN parameter estimation using prior distributions (system T5) also gives almost the same values for the warping factors and similar performance as using the lower order features, T6 system.

VTLN has a limited number of parameters (single warping factor in case of all-pass transform) to be estimated. On one hand, this enables estimation of warping factors and adaptation using very little adaptation data. On the other hand, there are only limited characteristics that this parameter can capture. Hence, in order to obtain improvements in adaptation when more adaptation data is available, VTLN may need to be combined with other adaptation methods such as CMLLR. This could result in additive improvements to CMLLR as shown in Saheer et al. [2010c]. There are studies in literature that show additive performance improvements with this combination and some others that state otherwise. This hypothesis will be tested in this work. The following chapter presents different techniques in order to combine VTLN with CMLLR and the research mainly focusses on implementing VTLN as a prior to CMLLR transform as in constrained structural maximum a posteriori linear regression (CSMAPLR).

5.4 Summary of Contributions

This work presents an efficient and accurate implementation of VTLN based on EM for the all-pass transform warping on MGCEP features. The EM implementation improves upon the grid search technique in time and space complexity. It also enables use of VTLN as a model transformation embedded into the HMM training and hence, works in similar ways to other model transformations like CMLLR/CSMAPLR. Appropriate warping factors for TTS are analyzed and techniques that were suggested to estimate similar values from the model are evaluated. Regression class based multiple transform VTLN is also presented which applies different warping factors to different state distributions. Subjective and Objective evaluations were performed on all the techniques proposed in the earlier chapter to work around the challenges in VTLN. The efficient way to estimate the warping factor for TTS is concluded as using lower order features for warping factor estimation and preferably avoiding extreme warping factor values using Jacobian normalization. The work presented in this chapter was published as Saheer et al. [2010a] and Saheer et al. [2012a].

6 Enhancing VTLN performance

VTLN as an effective rapid speaker adaptation technique was discussed in the previous chapters. There are a few drawbacks for VTLN due to its simple representation, that could be taken care of by viewing VTLN in a different perspective. Considering VTLN as a parametrically constrained version of feature space MLLTs (e.g. CMLLR) many possibilities emerge from the VTLN approach proposed in this work, in particular, as extensions of existing approaches in maximum likelihood linear transformation. VTLN could have an upper hand when the amount of adaptation data is small and pave the way to the powerful model adaptation techniques like CMLLR or CSMAPLR as and when more adaptation data comes in. To this end, this chapter explores various novel approaches to combine VTLN with CSMAPLR based model transformations.

6.1 Related Work

CMLLR was presented earlier as a powerful model adaptation technique. There are more robust linear transform based adaptation techniques that could prove beneficial in combination with VTLN. Structural maximum a posteriori (SMAP) based adaptation techniques proposed by Shiohan et al. [2002] uses prior information for transform estimation. The SMAP technique uses a family of elliptically symmetric distributions including the matrix variate normal prior density as a prior distribution [Chou, 1999] and uses a tree structure to propagate this prior to different classes of transforms. Yamagishi et al. [2009a] showed that due to the presence of a hierarchical prior, constrained SMAP linear regression (CSMAPLR) is a more robust adaptation framework when compared to CMLLR in the context of statistical parametric speech synthesis.

6.1.1 CSMAPLR

CMLLR is a model adaptation technique which can be shown to equivalently transform the spectral features ($\boldsymbol{x}$) as follows

$$\tilde{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{W}\boldsymbol{\xi} \tag{6.1}$$

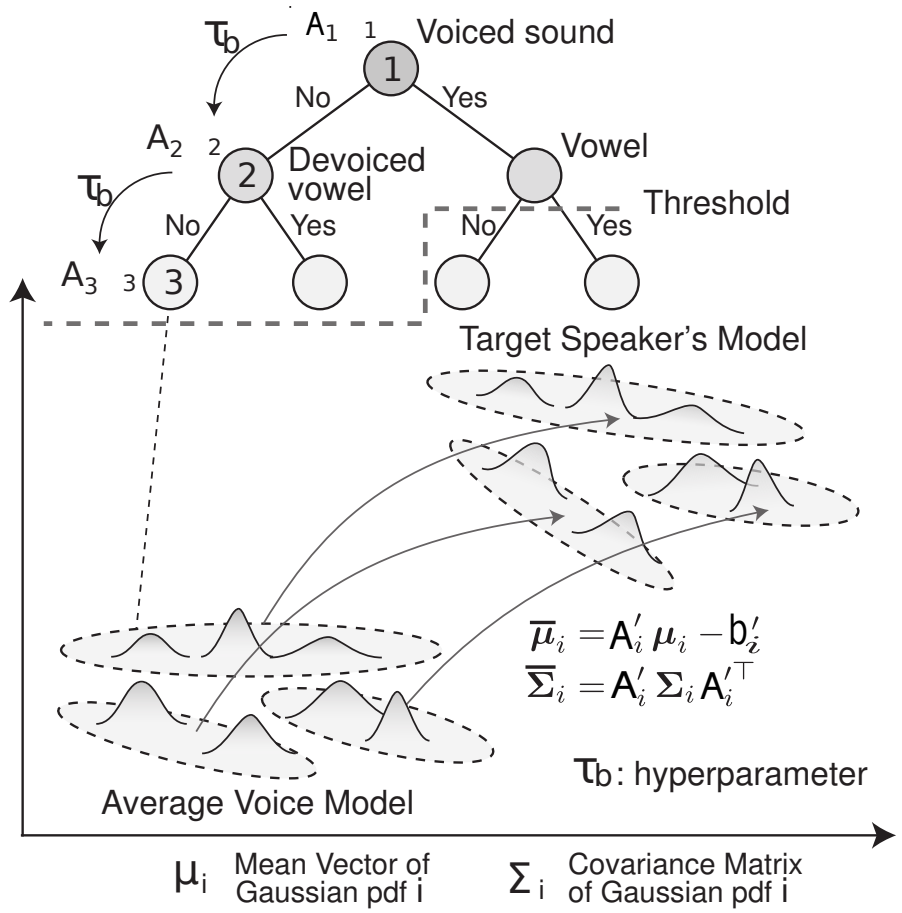


Figure 6.1 – CSMAPLR transformation based on Yamagishi et al. [2009a]

where $\boldsymbol{\xi} = [\boldsymbol{x}^\top, 1]^\top$, and $\boldsymbol{W} = [\boldsymbol{A}, \boldsymbol{b}]$. Note that, the transformation matrix $\boldsymbol{A}$ and bias vector $\boldsymbol{b}$ of the CMLLR transform are far less constrained than those for VTLN.

Constrained structural MAP based linear regression (CSMAPLR), shown in Figure 6.1, is a robust framework to estimate the CMLLR transforms $\boldsymbol{W}$ based on the SMAP criterion proposed by Shinoda and Lee [2001]:

$$\widehat{\boldsymbol{W}}_s = \underset{\boldsymbol{W}}{\operatorname{argmax}} P(\boldsymbol{x} \mid \Theta, \boldsymbol{W}_s, w_s)\, P(\boldsymbol{W}_s) \tag{6.2}$$

where $\boldsymbol{W}_s$ refers to the CMLLR transforms for the target speaker s. $P(\boldsymbol{x}_s \mid \Theta, \boldsymbol{W}_s, w_s)$ is a likelihood function for $\boldsymbol{W}_s$ and $P(\boldsymbol{W}_s)$ is a prior distribution of the transform $\boldsymbol{W}_s$. Θ represents the model parameters and w_s represents the transcriptions corresponding to the feature vector $\boldsymbol{x}$ for a speaker s. CMLLR transformations do not have conjugate prior distributions. There are not many options for the prior distribution of a matrix. Matrix variate normal distribution is probably the only viable distribution that can be used as the prior distribution $P(\boldsymbol{W})$ in this

case. The matrix variate normal distribution is given by:

$$P(\boldsymbol{W}) \propto |\boldsymbol{\Omega}|^{-\frac{L+1}{2}} |\boldsymbol{\Psi}|^{-\frac{L}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{W}-\boldsymbol{H})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{W}-\boldsymbol{H})\boldsymbol{\Psi}^{-1}\right] \tag{6.3}$$

where $\boldsymbol{\Omega} \in \mathbb{R}^{L\times L}$, $\boldsymbol{\Psi} \in \mathbb{R}^{(L+1)\times(L+1)}$ and $\boldsymbol{H} \in \mathbb{R}^{L\times(L+1)}$ are the hyperparameters of the prior distribution. There are a number of hyperparameters that need to be initialized for the matrix variate normal distribution. In the CSMAPLR estimation, the hyperparameter $\boldsymbol{\Psi}$ is fixed to the identity matrix and $\boldsymbol{\Omega}$ to a scaled identity matrix, $\boldsymbol{\Omega} = \tau_b \boldsymbol{I}_L$. τ_b is a positive scalar that controls the scale factor for the prior propagation and $\boldsymbol{I}_L$ is $L \times L$ identity matrix.

In the SMAP criterion, the tree structures of the distributions effectively control these hyperparameters. The whole adaptation data is used to estimate a global transform at the root node of the tree based on the ML criterion and it is propagated to the child nodes as a hyperparameter $\boldsymbol{H}$. At the root node the transform, $\boldsymbol{W}_1$ (shown in Figure 6.1) is estimated using the ML criterion as:

$$\boldsymbol{W}_1 = [\boldsymbol{A}_1, \boldsymbol{b}_1] = \underset{\boldsymbol{W}}{\mathrm{argmax}}\, P(\boldsymbol{x} \mid \Theta, \boldsymbol{W}, w_s) \tag{6.4}$$

The transforms at each child node are estimated using the corresponding adaptation data and hyperparameters propagated with the MAP criterion. The use of the prior distribution ensures that the transform at the node is restricted to be from a distribution centered around the transform of the parent node. The transform at the Node 2 in figure 6.1, represented as $\boldsymbol{W}_2$ depends on the transform $\boldsymbol{W}_1$ of the root node and is given by:

$$\boldsymbol{W}_2 = [\boldsymbol{A}_2, \boldsymbol{b}_2] = \underset{\boldsymbol{W}}{\mathrm{argmax}}\, P(\boldsymbol{x} \mid \Theta, \boldsymbol{W}, w_s)|\tau_b|^{-\frac{L+1}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{W}-\boldsymbol{A}_1)^{\top}[\tau_b \boldsymbol{I}_L]^{-1}(\boldsymbol{W}-\boldsymbol{A}_1)\right] \tag{6.5}$$

The transform at Node 2 is used as the mean of the prior distribution at Node 3 and the transform $\boldsymbol{W}_3$ at this node is given by:

$$\boldsymbol{W}_3 = [\boldsymbol{A}_3, \boldsymbol{b}_3] = \underset{\boldsymbol{W}}{\mathrm{argmax}}\, P(\boldsymbol{x} \mid \Theta, \boldsymbol{W}, w_s)|\tau_b|^{-\frac{L+1}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{W}-\boldsymbol{A}_2)^{\top}[\tau_b \boldsymbol{I}_L]^{-1}(\boldsymbol{W}-\boldsymbol{A}_2)\right] \tag{6.6}$$

This process is continued recursively from the root node to all the leaf nodes of the tree structure. The tree structure used is based on linguistic information and the propagated prior distribution thus reflects the connection and similarity between the distributions of linguistic information. The hyperparameter of the prior distribution $\boldsymbol{H}$ at the root node of the tree structure could be set to an identity matrix, that is, a prior favouring no occupancy smoothing.

The re-estimation formulas based on the Baum-Welch algorithm for the transformation matrix is given by:

$$\hat{\boldsymbol{w}}_l = (\kappa \boldsymbol{p}_l + \boldsymbol{k}_l){\boldsymbol{G}_l}^{-1} \tag{6.7}$$

where $\hat{\boldsymbol{w}}_l$ represents the l^{th} row of the transformation matrix with $\boldsymbol{p}_l = [0, \boldsymbol{c}_l]$, $\boldsymbol{c}_l$ is the l^{th} cofactor row vector of the transformation matrix, $\boldsymbol{W}$. The value κ satisfies the quadratic equation:

$$\kappa^2 \boldsymbol{p}_l \boldsymbol{G}_l^{-1} (\boldsymbol{p}_l)^T + \kappa \boldsymbol{p}_l \boldsymbol{G}_l^{-1} (\boldsymbol{k}_l)^T - \sum_{m=1}^{M} \sum_{f=1}^{F} \gamma_{m.f} = 0 \tag{6.8}$$

where M is the total number of mixtures, F is the total number of frames and $\gamma_{m,f}$ is the posterior probability of mixture, m at frame, f. The MAP representation is reflected in the $\boldsymbol{k}_l$ and $\boldsymbol{G}_l$ parameters as shown by Yamagishi et al. [2009a].

$$\boldsymbol{G}_l = \sum_{m=1}^{M} \frac{1}{\boldsymbol{\sigma}_{m_l}^2} \sum_{f=1}^{F} \boldsymbol{\gamma}_{m,f} \boldsymbol{\xi}_f \boldsymbol{\xi}_f^T + \tau_b \boldsymbol{I}_L \tag{6.9}$$

$$\boldsymbol{k}_l = \sum_{m=1}^{M} \frac{1}{\boldsymbol{\sigma}_{m_l}^2} \boldsymbol{\mu}_{m_i} \sum_{f=1}^{F} \boldsymbol{\gamma}_{m,f} \boldsymbol{\xi}_f^T + \tau_b \boldsymbol{H}_l \tag{6.10}$$

where $\boldsymbol{H}_l$ is the l^{th} row of the matrix $\boldsymbol{H}$.

6.1.2 VTLN with linear transforms

Since it is known that VTLN can capture very few speaker characteristics, there have been many attempts to combine it with other linear transformations. One of the first attempts could be by Pye and Woodland [1997] to combine VTLN with MLLR transforms for speaker adaptive training. It was shown to give additive performance. There were some other studies [Uebel and Woodland, 1999] that proposed VTLN with multiple iterations of CMLLR may not give additive improvements unlike with unconstrained MLLR. It was observed that after 6 iterations of CMLLR, there was no significant advantage in adding VTLN. This was postulated to the fact that CMLLR is performing very similar operation compared to linear VTLN and is more powerful in having more transformation variables and estimates separate transforms for static and differential cepstral coefficients, VTLN can be even viewed as a more constrained form of CMLLR, constrained on the vocal tract length. It was mentioned by Uebel and Woodland [1999] that estimating both transforms would be no better than just using CMLLR unless the effect of initialization is of key importance which is the case for not having additional performance improvements after multiple iterations of CMLLR.

It was shown by Panchapagesan and Alwan [2009] that estimating a bias vector and unconstrained variance transformation according to MLLR on top of the linear transform based frequency warping can further improve the recognition accuracy. This phenomenon is mainly observed with very limited adaptation data of the order of one adaptation sentence compared to the MLLR transforms which outperform with more adaptation data.

Breslin et al. [2010] showed that VTLN can be combined with constrained MLLR (CMLLR) for rapid adaptation in ASR. In that work, a count smoothing framework is used to incorporate the prior information. The count smoothing framework was initially presented by Flego and Gales [2009], where the predictive and adaptive noise compensating transforms were combined using this scheme. The predictive approaches make use of a mismatch function that represents the impact of the background noise on the clean speech. The number of parameters associated with this mismatch function is usually small. This is in contrast to adaptive approaches to speaker and noise compensation where, normally, linear transforms of the model parameters are estimated. A large number of transforms and associated parameters must be estimated. Flego and Gales [2009] mention that CMLLR does not have a conjugate prior, instead count smoothing can be used to combine it with the predictive transforms. The pseudo counts associated with the predictive transform are combined with the actual observed counts and the transforms are estimated. Breslin et al. [2010] used this count smoothing framework to combine rapid adaptation techniques like VTLN and predictive CMLLR (pCMLLR) with CMLLR adaptive transforms. Thus, using a dynamic prior instead of a static prior. Statistics for estimating the CMLLR transform, $\boldsymbol{G}_i$ and $\boldsymbol{k}_i$, are based on interpolating adaptive and prior statistics, given by

$$\boldsymbol{G}_i = \boldsymbol{G}_{adi} + \tau \frac{\boldsymbol{G}_{pri}}{\sum_m \gamma_m} \tag{6.11}$$

$$\boldsymbol{k}_i = \boldsymbol{k}_{adi} + \tau \frac{\boldsymbol{k}_{pri}}{\sum_m \gamma_m} \tag{6.12}$$

The prior statistics $\boldsymbol{G}_{pri}$ and $\boldsymbol{k}_{pri}$ are normalized so that they effectively contribute τ frames to the final statistics. γ_m represents the occupancy count for the output distributions. As more data becomes available, the adaptive CMLLR statistics $\boldsymbol{G}_{adi}$ and $\boldsymbol{k}_{adi}$ will dominate, but for small amounts of data the prior statistics are more important. VTLN was also tested as a parent transform for CMLLR representing a cascade of transforms. It was shown that using VTLN in the direct transform estimation framework as a prior for estimating CMLLR transforms was more robust than standard CMLLR to small amounts of data, and more robust than using a cascade of transforms where no prior was used. It was shown by Karhila et al. [2012] that stacked VTLN and CSMAPLR transformations can improve child speech synthesis. The adult average voice was transformed using child speech using speaker adaptation and used to synthesize specific child voices using speaker transformations. VTLN and CSMAPLR transformations are used as parent and child transformations to get the effect of a cascade.

6.2 Combining VTLN with linear transforms

Maximum likelihood linear transformation (MLLT) based adaptation techniques entail linear transformation of the means and variances of an HMM to match the characteristics of the

speech for a given speaker. These techniques require a considerable amount of adaptation data (of the order of tens of utterances) for reasonable adaptation performance. By contrast, a rapid adaptation technique like VTLN requires very little adaptation data as it estimates only a single parameter. This system preserves the naturalness of the average voice, albeit capturing limited speaker characteristics. It follows that combining the linear transform based adaptation techniques with VTLN could result in improved naturalness of synthesized speech whilst also being effective at capturing the speaker characteristics. This provides a means to rapidly adapt synthesized speech with a balanced trade-off between naturalness and speaker similarity. This section presents different techniques to combine VTLN with the model based transformations.

6.2.1 VTLN as prior for CSMAPLR

CSMAPLR uses ML estimation at the root node and hence, there is no prior distribution at the root node. The MAP criterion can be used at the root node as well and the hyper parameter at the root node may be replaced by a VTLN transform. The structural framework helps propagate the prior information affected by the VTLN transform through the various levels of the regression tree effectively. The tree structure is generated using linguistic information; hence, the propagated prior information should reflect the connection and similarity of the distributions of linguistic information. Using the VTLN matrix as the initial prior information for the root node of the CSMAPLR transform could result in the propagation of speaker characteristics and improved speaker adaptation even when very little data is available.

The VTLN transformation presented in this work can be considered as a very constrained form of CMLLR/CSMAPLR. The single parameter normally gives some measure of the vocal tract length, but more concretely is known to be highly correlated with basic speaker characteristics such as gender and as such can act as the mean of the prior distribution for speaker independent modeling. Usually, CSMAPLR uses all the data to generate the transformation at the root node and uses a maximum likelihood (ML) estimation at the root node without using any prior distribution. In fact the CSMAPLR adaptation technique can use any arbitrary prior information (even the identity matrix) at the root node of the tree structure. This prior information can easily be replaced with the VTLN transformation matrix (refer Figure 6.2). At the root node, we may set the hyperparameter $\boldsymbol{H}$ representing the mean of the prior distribution as

$$\boldsymbol{H}_{\text{VTLN}} = [\boldsymbol{A}_\alpha, \mathbf{0}] \tag{6.13}$$

where $\boldsymbol{A}_\alpha$ is the VTLN transformation matrix described by α and $\mathbf{0}$ is a zero bias vector. The VTLN prior may be used for the dynamic features of the cepstra; in this case the hyperparame-

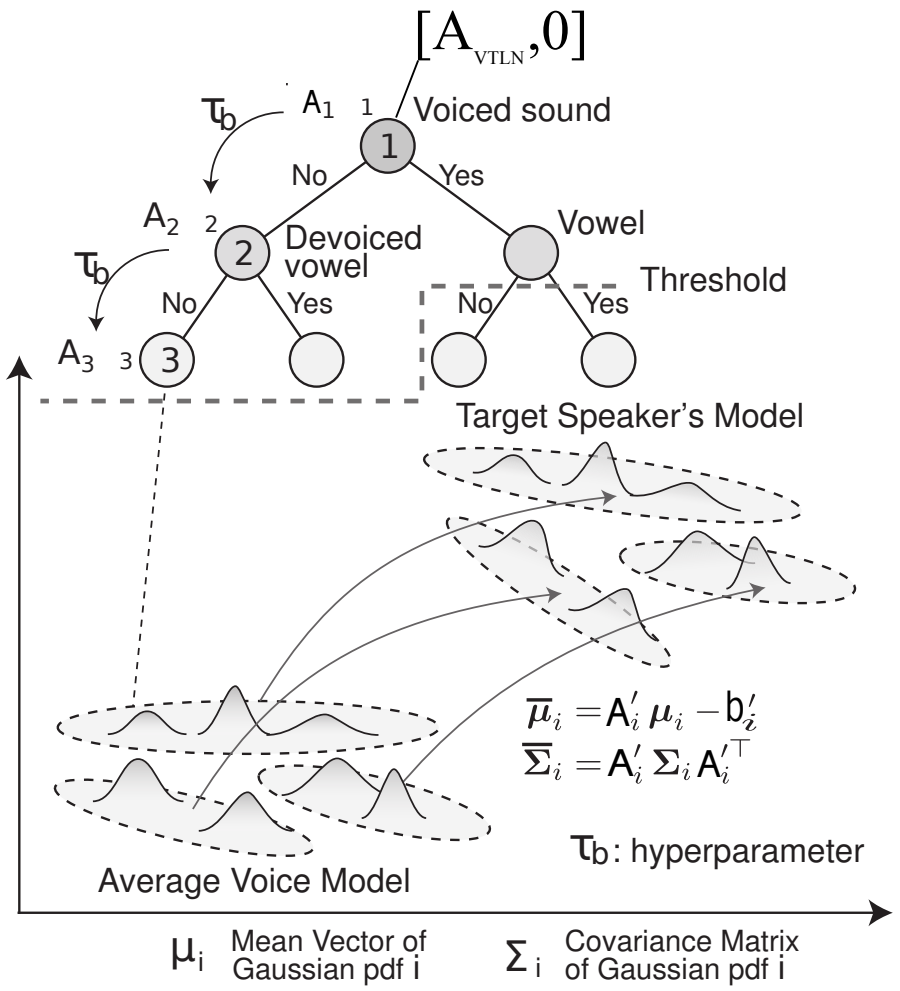


Figure 6.2 – VTLN transformation matrix is used as prior for the root node of the CSMAPLR transformation.

ter matrix $\boldsymbol{H}$ is a block diagonal matrix with repeating $\boldsymbol{A}_\alpha$ matrix.

$$\boldsymbol{H}_{VTLN} = \begin{bmatrix} \boldsymbol{A}_\alpha & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_\alpha & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_\alpha & \boldsymbol{0} \end{bmatrix} \tag{6.14}$$

While propagating the prior information through the lower nodes of the tree, τ_b is the scale factor determining the influence of the VTLN prior on the CSMAPLR adaptation technique. τ_b manipulates the variance of the prior distribution and hence determines the effect of the prior in the MAP estimation. The value of the scale factor can be empirically estimated depending on the availability of adaptation data. Scale factors in the range of 1 to 10000 are used to generate adaptation transforms and objective (MCD) score is used as the metric to determine the apt value.

Following similar equations as the Baum-Welch re-estimation for the CSMAPLR transformation matrix, the parameters at the root node of the VTLN+CSMAPLR estimation is given by:

$$\boldsymbol{G}_l = \sum_{m=1}^{M} \frac{1}{\boldsymbol{\sigma}_{m_l}^2} \sum_{f=1}^{F} \boldsymbol{\gamma}_{m,f} \boldsymbol{\xi}_f \boldsymbol{\xi}_f^T + \tau_b \boldsymbol{I}_L \tag{6.15}$$

$$\boldsymbol{k}_l = \sum_{m=1}^{M} \frac{1}{\boldsymbol{\sigma}_{m_l}^2} \boldsymbol{\mu}_{m_i} \sum_{f=1}^{F} \boldsymbol{\gamma}_{m,f} \boldsymbol{\xi}_f^T + \tau_b \boldsymbol{A}_{\alpha,l} \tag{6.16}$$

where $\boldsymbol{A}_{\alpha,l}$ is the l^{th} row of the VTLN transformation matrix $\boldsymbol{A}_\alpha$.

The gender characteristics estimated by VTLN when propagated to the nodes of the tree structure are expected to improve the speaker specific transform estimation for CSMAPLR. More specifically, VTLN has been shown to be closer to the average voice, and hence better in naturalness [Saheer et al., 2010a] and CSMAPLR is known to bring in better speaker similarity when very little adaptation data is available. A-priori, combination of these two is expected to give improved performance with respect to naturalness and speaker similarity.

6.2.2 Tree structure

The maximum a posteriori linear regression (MAPLR) as such is a powerful adaptation technique. Though CSMAPLR gives a structure to the transform estimation, it would be interesting to check whether the gains obtained from the VTLN prior is due to the tree structure and the hierarchical transform estimation or just the effect of VTLN as such as the prior. To this end, it is possible to use VTLN prior for each node of the regression tree which otherwise uses a hierarchical prior. This results in a MAPLR implementation with VTLN as the mean of the prior distribution instead of using CSMAPLR as the principal transformation. A global VTLN transform could be used or multiple VTLN transforms could be used as the prior for different nodes of the regression tree.

6.2.3 Stacked Transforms

The idea of stacked transformations is to use multiple transforms where each transform has a different role to play. Similar to investigations performed by Breslin et al. [2010] and Karhila et al. [2012], a cascade of transforms can be used to combine VTLN with other linear transformations like CSMAPLR. Initially, a VTLN linear transformation can be estimated and then use this transform as the parent transform to estimate the CSMAPLR transformations. VTLN can represent the gender of the speaker and then CSMAPLR can capture more detailed speaker characteristics. Thus enabling a cascade of transformations could be an effective way of combining VTLN with CSMAPLR. The stacked transforms is similar to the prior technique explained in the previous section. But, it becomes useful if the different sources of variations

(e.g. speaker and environment) are to be factored out. In this case, both levels of transforms in the stacked arrangement are taking into account speaker information and VTLN is acting as a quasi-prior (at least in the E-step). As mentioned earlier by Breslin et al. [2010], the importance of using stacked transform is that it provides the possibility to converge to a better estimate of CMLLR transform since the first transform in the cascade (VTLN) leads to a better estimate of the hidden variables in the E-step. Whether this benefit would be maintained after multiple iterations needs to be investigated. Also, it should be interesting to check out if the combination can really perform well with very little adaptation data and be useful as a rapid adaptation technique as claimed by Karhila et al. [2012].

6.3 Bias for VTLN

Similar to other linear transformations, VTLN is implemented as a linear transformation of the cepstrum or equivalently model parameters. The linear transformations have two important components, viz. translation and scaling represented as $\boldsymbol{W} = [\boldsymbol{A}, \boldsymbol{b}]$. Where, A represents the scaling and b represents translation usually referred to as bias.

Bias is a very important term in adaptation using linear transformations and influences the performance a lot. It was shown by Lööf et al. [2008] that a set of offset transforms alone without any scaling/rotation for the mean of the Gaussian models, termed as shift-MLLR, could be used for generating better speaker adaptive models. The offset terms are the bias terms of the speaker transformation. Current implementations of VTLN do not estimate a bias term. The derivation of the bias term for VTLN is shown below. Unlike for the scaling and rotation matrix $\boldsymbol{A}$, derivation of the bias term is not exactly similar for CMLLR and VTLN.

The auxiliary function to optimize using the maximum likelihood (ML) technique is given by:

$$Q = \log \mathscr{L}(\boldsymbol{x}(t); \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}_\alpha, \boldsymbol{b}) \tag{6.17}$$

where, L represents the likelihood function, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ represent the mean and variance of the model with $\boldsymbol{x}$ as the feature vector. $\boldsymbol{A}_\alpha$ represents the VTLN transformation matrix and $\boldsymbol{b}$, the bias term for VTLN to be derived.

Using a mixture of Gaussians for the state probability distributions with γ_m as the occupancy for each mixture,m yields the following equation with a Jacobian term given by $\log|\boldsymbol{A}_\alpha|$.

$$Q = \sum_t \sum_m \gamma_m \Big[\log(N(\boldsymbol{A}_\alpha x_t + b \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)) + \log|\boldsymbol{A}_\alpha| \Big] \tag{6.18}$$

$$= \sum_t \sum_m \gamma_m \Big[\log\Big(\frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma_m}|}} \times \exp\Big\{ -\frac{1}{2}(\boldsymbol{A}_\alpha \boldsymbol{x}_t + \boldsymbol{b} - \boldsymbol{\mu}_m)^T {\boldsymbol{\Sigma}_m}^{-1} (\boldsymbol{A}_\alpha \boldsymbol{x}_t + \boldsymbol{b} - \boldsymbol{\mu}_m) \Big\} \Big) + \log|\boldsymbol{A}_\alpha| \Big] \tag{6.19}$$

Applying log and ignoring constants,

$$Q = \sum_t \sum_m \gamma_m \Big[-\frac{1}{2}\log|\boldsymbol{\Sigma}_m| - \frac{1}{2}(\boldsymbol{A}_\alpha \boldsymbol{x}_t + \boldsymbol{b} - \boldsymbol{\mu}_m)^T {\boldsymbol{\Sigma}_m}^{-1}(\boldsymbol{A}_\alpha \boldsymbol{x}_t + \boldsymbol{b} - \boldsymbol{\mu}_m) + \log|\boldsymbol{A}_\alpha| \Big] \quad (6.20)$$

Ignoring the terms independent of b results in

$$Q = \sum_t \sum_m \gamma_m \Big[-\frac{1}{2}(\boldsymbol{A}_\alpha \boldsymbol{x}_t + \boldsymbol{b} - \boldsymbol{\mu}_m)^T {\boldsymbol{\Sigma}_m}^{-1}(\boldsymbol{A}_\alpha \boldsymbol{x}_t + \boldsymbol{b} - \boldsymbol{\mu}_m) \Big] \quad (6.21)$$

To optimize this auxiliary function using expectation maximization (EM), calculate the derivative of this function on bias 'b'. Following a similar derivation for mean in Garner and Holmes [1998] and Liporace [1982].

$$Q = -\frac{1}{2}\sum_t \sum_m \gamma_m \Big[(\boldsymbol{A}_\alpha \boldsymbol{x}_t + \boldsymbol{b} - \boldsymbol{\mu}_m)^T {\boldsymbol{\Sigma}_m}^{-1}(\boldsymbol{A}_\alpha \boldsymbol{x}_t + \boldsymbol{b} - \boldsymbol{\mu}_m) \Big] \quad (6.22)$$

Using the standard matrix quadratic differential calculus formula:

$$\frac{\partial}{\partial x}(\boldsymbol{y} - \boldsymbol{x})^T \boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x}) = -\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x}) - \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{x}) \quad (6.23)$$

$$\frac{\partial Q}{\partial b} = \frac{1}{2}\sum_t \sum_m \gamma_m \Big[\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{A}_\alpha \boldsymbol{x}_t - \boldsymbol{\mu}_m + \boldsymbol{b}) + {\boldsymbol{\Sigma}_m^{-1}}^T(\boldsymbol{A}_\alpha \boldsymbol{x}_t - \boldsymbol{\mu}_m + \boldsymbol{b}) \Big] \quad (6.24)$$

$$= \sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{A}_\alpha \boldsymbol{x}_t - \boldsymbol{\mu}_m + \boldsymbol{b}) \quad (6.25)$$

Equating the RHS to zero to find the maximum

$$\frac{\partial Q}{\partial b} = \sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{A}_\alpha \boldsymbol{x}_t - \boldsymbol{\mu}_m + \boldsymbol{b}) = 0 \quad (6.26)$$

$$-\sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{A}_\alpha \boldsymbol{x}_t - \boldsymbol{\mu}_m) = \sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1} \boldsymbol{b} \quad (6.27)$$

Multiplying both sides by the inverse of the statistics over the inverse of the covariance ($\boldsymbol{\Sigma}_m^{-1}$):

$$b = -\Big(\sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1}\Big)^{-1} \sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{A}_\alpha \boldsymbol{x}_t - \boldsymbol{\mu}_m) \quad (6.28)$$

Using a diagonal covariance matrix:

$$b = -\Big(\sum_t \sum_m \gamma_m \sum_i \frac{1}{\boldsymbol{\sigma}_{m,i}^2}\Big)^{-1} \sum_t \sum_m \gamma_m \sum_i \frac{(\boldsymbol{A}_{\alpha,i} \boldsymbol{x}_{t,i} - \boldsymbol{\mu}_{m,i})}{\boldsymbol{\sigma}_{m,i}^2} \quad (6.29)$$

The resulting bias term could be implemented with the linear VTLN transformation. Since it is not possible to estimate the transformation matrix and bias term simultaneously, a better

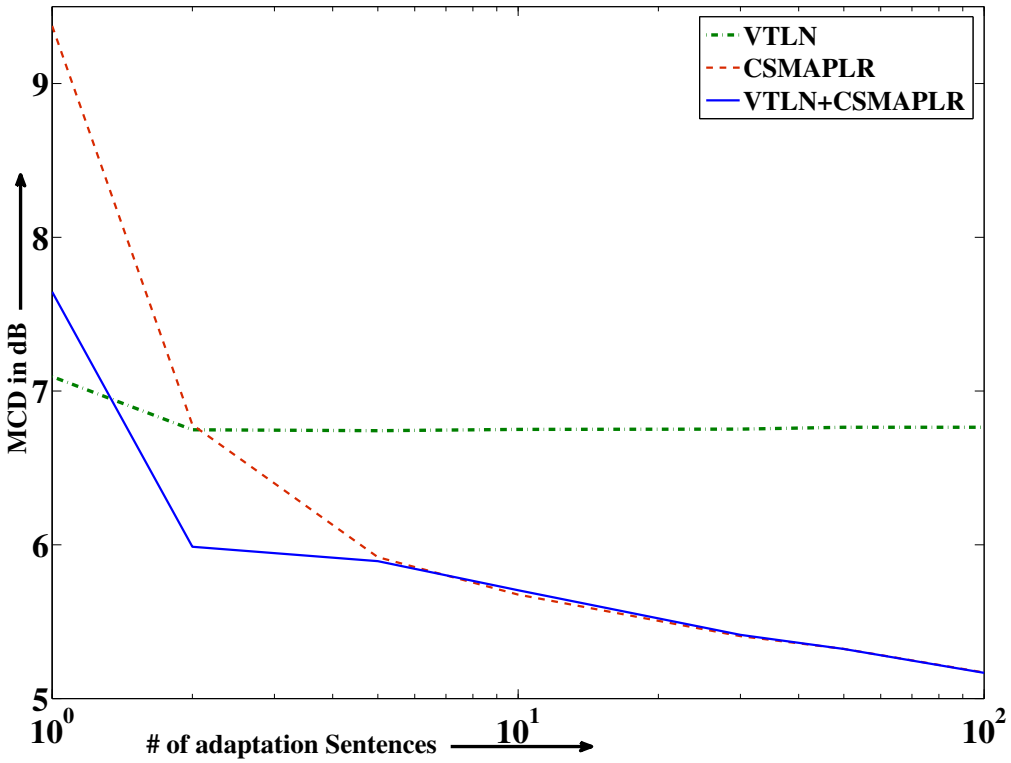


Figure 6.3 – MCD for VTLN, CSMAPLR and the proposed VTLN-CSMAPLR.

approach would be to iteratively optimize the two components each estimated in alternative iterations considering the effect of the other term.

6.4 Experiments & Results

A number of different techniques were presented in the previous section to combine the advantages of VTLN and more powerful CSMAPLR transformations. Such a combination takes the advantage of both techniques. When the adaptation data is scarce, VTLN will improve the quality and as more data comes in CSMAPLR is expected to bring in better speaker characteristics. This section presents experiments to evaluate the proposed combinations. These techniques are independent of the underlying models and could be used effectively in ASR or TTS. Hence, a few techniques are evaluated using an ASR system which is easier to evaluate compared to (subjective evaluations of) TTS. This also adds further evidence to the unification (of adaptation techniques in ASR and TTS) theme of this thesis.

6.4.1 TTS

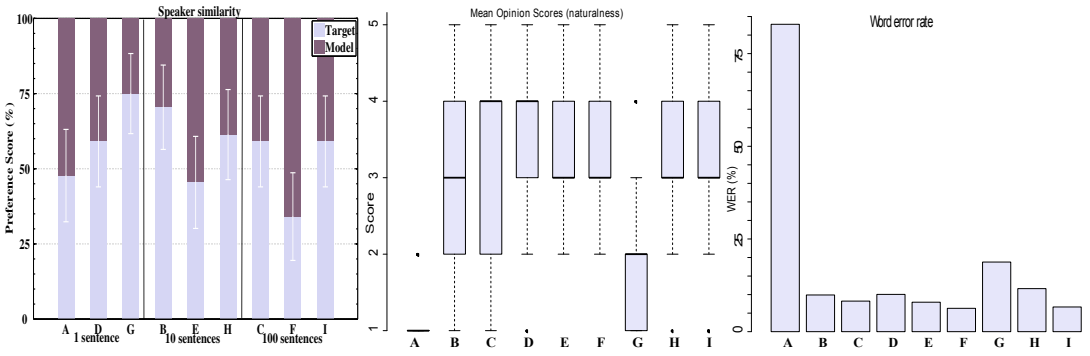
The HMM speech synthesis system (HTS) by Zen et al. [2009] was used for generating the statistical parameters for speech synthesis. HTS models spectrum, $\log F_0$, band-limited aperiodic components and duration in the unified framework of hidden semi-Markov models (HSMMs). The STRAIGHT vocoder was used to synthesize speech from the parameters gener-

ated using HTS. The HMM topology was five-state and left-to-right with no skip states. Speech was recorded at 48kHz sampling rate and the features were 59th-order mel-cepstra, $\log F_0$, 25-dimensional band aperiodicity, and their delta and delta-delta coefficients, extracted from 48kHz recordings with a frame shift of 5ms. The speaker dependent model was built using a UK English speech corpus including 5 hours of clean speech data uttered by an RP professional narrator. The evaluation experiments were performed on another UK English test speaker. Subjective listening tests were performed in two parts each comparing different amounts of adaptation data, 11 and 40 subjects participated in the first and second subjective evaluations respectively. The speech to be evaluated was generated using the Blizzard challenge 2010 test sentences and tested for naturalness, speaker similarity and intelligibility with different amounts of adaptation data and different values of the scale factor.

The subjective tests were based on mean opinion scores (MOS) of naturalness and speaker similarity and additional ABX scores for speaker similarity. The synthesized utterances were rated on a 5-point scale for the MOS tests, 5 being "completely natural" or "sounds exactly like the target speaker" and 1 being "completely unnatural" or "sounds like a totally different speaker". In the ABX test, listeners were presented with a test utterance (X) and two reference voices (A and B) and asked to select the reference that sounded closest to the test utterance. The model (speaker used to train the model) and the target speaker were given as the two reference speakers in the ABX test for finding speaker similarity. Only the spectral stream was transformed with different adaptation techniques; other streams (logF0, bndap and duration) were unadapted or the same as generated for the speaker used to train the model. The subjective evaluations were also performed for intelligibility using semantically unpredictable sentences where subjects listen to the speech utterances and were asked to type the corresponding text. The score for intelligibility was based on the word error rate (WER) for the text entered by the listeners. In addition, objective evaluation based on the mel-cepstral distance (MCD) was also carried out. The MCD is the Euclidean distance between the synthesized cepstra and those derived from the natural speech, and can be viewed as an approximation to the log spectral distortion measure according to Parserval's theorem. One hundred sentences were synthesized for objective evaluations for the test speaker.

Results and Discussion

The values of the MCD scores for different amounts of adaptation data are plotted in the Figure 6.3. The figure shows the MCD score for the scale factor of 1000 (which was empirically determined to be appropriate) for both CSMAPLR and VTLN+CSMAPLR. The objective results show that 1) the VTLN technique works best in comparison to others when one adaptation sentence is used (around 7dB) whereas its performance does not improve if more than one sentence is used for the adaptation and that 2) the CSMAPLR improves the MCD to around 6dB when the number of adaptation sentences is more than five. However, the performance of the CSMAPLR technique rapidly becomes worse when the number of adaptation sentences is less than five, reaching around 9.5dB MCD with only one adaptation utterance. Finally,



System symbol Correspondence

System	Adaptation	Adp #
A	CSMAPLR	1
B	CSMAPLR	10
C	CSMAPLR	100
D	VTLN	1
E	VTLN	10
F	VTLN	100
G	VTLN-CSMAPLR	1
H	VTLN-CSMAPLR	10
I	VTLN-CSMAPLR	100

Significance at 1% level for **similarity**

	A	B	C	D	E	F	G	H	I
A	-	0	0	0	0	0	1	0	0
B	0	-	0	0	0	1	0	0	0
C	0	0	-	0	0	0	0	0	0
D	0	0	0	-	0	0	0	0	0
E	0	0	0	0	-	0	1	0	0
F	0	1	0	0	0	-	1	0	0
G	1	0	0	0	1	1	-	0	0
H	0	0	0	0	0	0	0	-	0
I	0	0	0	0	0	0	0	0	-

Significance at 1% level for **naturalness**

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	1	1	1	1	1
B	1	-	0	0	0	0	1	0	0
C	1	0	-	0	0	0	1	0	0
D	1	0	0	-	0	0	1	0	0
E	1	0	0	0	-	0	1	0	0
F	1	0	0	0	0	-	1	0	0
G	1	1	1	1	1	1	-	1	1
H	1	0	0	0	0	0	1	-	0
I	1	0	0	0	0	0	1	0	-

Significance at 1% level for **WER**

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	1	1	1	1	1
B	1	-	0	0	0	0	0	0	0
C	1	0	-	0	0	0	0	0	0
D	1	0	0	-	0	0	0	0	0
E	1	0	0	0	-	0	0	0	0
F	1	0	0	0	0	-	0	0	0
G	1	0	0	0	0	0	-	0	0
H	1	0	0	0	0	0	0	-	0
I	1	0	0	0	0	0	0	0	-

Figure 6.4 – Listening tests results. There are three columns of plots and tables which are, from left to right, similarity to original speaker, mean opinion score for naturalness, and intelligibility. The similarity is an ABX plot with whiskers for 95% confidence interval. Here systems are permuted differently for readability. Naturalness plot on the upper row is a box plot where the median is represented by a solid bar across a box showing the quartiles and whiskers extend to 1.5 times the inter-quartile range. The system-symbol correspondence is shown in the first table in the bottom row. The rest of the tables in the bottom row indicate significant differences between pairs of systems, based on Wilcoxon signed rank tests with alpha Bonferoni correction (1% level); '1' indicates a significant difference.

the objective results clearly show that the proposed VTLN-CSMAPLR technique alleviates this issue of the CSMAPLR technique and improves the performance when the number of adaptation sentences is less than five. We can see that even if the number of adaptation sentences is just two, the performance of the VTLN-CSMAPLR technique outperforms the VTLN technique; its distortion is around 6dB.

The first listening test was performed with 1, 10 and 100 adaptation sentences. The evaluation results of the listening tests are shown in Figure 6.4. From the speaker similarity results, we can see that VTLN works best when the number of adaptation sentences is one and also that VTLN-CSMAPLR outperforms CSMAPLR with one adaptation sentence. There is no significant difference among the CSMAPLR and VTLN-CSMAPLR adaptation methods with 10 or 100 adaptation sentences, both outperform the VTLN adaptation. From the results on naturalness, It can be seen that VTLN does not improve naturalness even if more data is used. However,

VTLN and VTLN-CSMAPLR both give better results than CSMAPLR with one adaptation sentence. From the intelligibility evaluation, it is observe that there is no significant difference between VTLN and VTLN-CSMAPLR with 1, 10 and 100 sentences, but, on the other hand, we can see that CSMAPLR has significantly degraded intelligibility with one adaptation sentence.

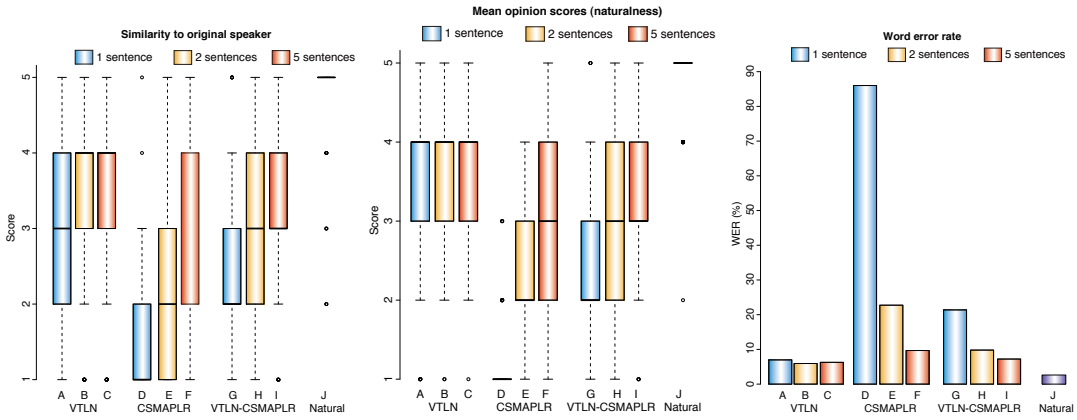
Two sets of listening tests were performed with different amounts of adaptation data. The first listening test confirmed that VTLN-CSMAPLR scales up to the performance of CSMAPLR with large amounts of adaptation data. Since, the 10 and 100 adaptation sentences seems to be large enough to show any perceivable difference in performance for CSMAPLR and VTLN-CSMAPLR, a second listening test with very little adaptation data could show the exact threshold on adaptation data for improvements with VTLN. The second listening test was performed with 1, 2 and 5 adaptation sentences. The evaluation results of the listening tests are shown in Figure 6.5. From the speaker similarity results, it can be seen that (as observed before), VTLN works best when the number of adaptation sentences is one or two and also that VTLN-CSMAPLR outperforms CSMAPLR with one or two adaptation sentences. There is no significant difference among the adaptation methods with five adaptation sentences. From the results on naturalness, we see that VTLN does not improve naturalness even if more data is used. However, VTLN and VTLN-CSMAPLR both give better results for naturalness than CSMAPLR with one or two adaptation sentences. From the intelligibility evaluation, it is observed again that there is no significant difference between VTLN and VTLN-CSMAPLR with two and five sentences, but, on the other hand, it can be seen that CSMAPLR has significantly degraded intelligibility with one adaptation sentence.

In both these results, VTLN is preferred even for speaker similarity only because the test speaker is very close to the speaker used to generate the speaker dependent model (both are RP English male speakers) and VTLN was much better in naturalness. With target speakers very different from the model speaker, the speaker similarity of VTLN will be very poor compared to that of VTLN-CSMAPLR or CSMAPLR. It can be concluded that the VTLN prior can significantly improve the CSMAPLR adaptation performance when the adaptation data is very limited and unlike VTLN, can scale up to the performance of CSMAPLR with more adaptation data. A more detailed evaluation of the different methods proposed in this chapter are presented using some special evaluations in chapter 7.

6.4.2 ASR

Following the work by Dines et al. [2009b,a], this section presents some ASR experiments to show that the techniques developed in this research can in fact be used for both TTS and ASR equally well and the work is in accordance to the unification theme of this thesis. It should be noted that the results may not be the state of the art for this database since the idea is to use a common model for ASR and TTS.

The hidden Markov models were built with 13 dimensional cepstral features with Δ and Δ^2 for the (US English) WSJ0 database. The models were built using single component mixture



Significance at 1% level for **speaker similarity**

	A	B	C	D	E	F	G	H	I	J
A	-	1	0	1	1	1	1	0	0	1
B	1	-	0	1	1	0	1	1	0	1
C	0	0	-	1	1	0	1	1	0	1
D	1	1	1	-	1	1	1	1	1	1
E	1	1	1	1	-	1	0	1	1	1
F	0	0	0	1	1	-	1	1	0	1
G	1	1	1	1	0	1	-	0	1	1
H	0	1	1	1	1	1	0	-	1	1
I	0	0	0	1	1	0	1	1	-	1
J	1	1	1	1	1	1	1	1	1	-

Significance at 1% level for **naturalness**

	A	B	C	D	E	F	G	H	I	J
A	-	0	0	1	1	1	1	1	1	1
B	0	-	0	1	1	1	1	1	1	1
C	0	0	-	1	1	1	1	1	1	1
D	1	1	1	-	1	1	1	1	1	1
E	1	1	1	1	-	1	0	1	1	1
F	1	1	1	1	1	-	1	0	0	1
G	1	1	1	1	0	1	-	1	1	1
H	1	1	1	1	1	0	1	-	1	1
I	1	1	1	1	1	0	1	1	-	1
J	1	1	1	1	1	1	1	1	1	-

Significance at 1% level for **word error rates**

	A	B	C	D	E	F	G	H	I	J
A	-	0	0	1	1	0	1	0	0	1
B	0	-	0	1	1	0	1	0	0	1
C	0	0	-	1	1	0	1	0	0	0
D	1	1	1	-	1	1	1	1	1	1
E	1	1	1	1	-	1	0	1	1	1
F	0	0	0	1	1	-	1	0	0	1
G	1	1	1	1	0	1	-	1	1	1
H	0	0	0	1	1	0	1	-	0	1
I	0	0	0	1	1	0	1	0	-	1
J	1	1	0	1	1	1	1	1	1	-

Figure 6.5 – Listening tests results. There are three columns of plots and tables which are, from left to right, similarity to original speaker, mean opinion score for naturalness, and intelligibility. The similarity and naturalness plots on the upper row are box plots where the median is represented by a solid bar across a box showing the quartiles and whiskers extend to 1.5 times the inter-quartile range. The tables in the bottom row indicate significant differences between pairs of systems, based on Wilcoxon signed rank tests with alpha Bonferoni correction (1% level); '1' indicates a significant difference.

PDFs to demonstrate the maximum impact of VTLN, and because only single component mixture models can be used in synthesis. In particular, this was to avoid the situation where multi-mixture models either over-fitted, or modeled the speaker variabilities that could be attributed to VTLN. The setup is very similar to the synthesis experiments and uses the HTS software to perform the evaluations. The single stream spectral features were extracted using STRAIGHT spectral analysis [Kawahara et al., 1999]. Speech recognition and synthesis systems use the same average voice training procedure which involves the generation of maximum likelihood speaker adaptive trained (SAT), context dependent, left-right models.

The ASR experiments were performed to compare the effect of VTLN prior on the recognition performance. CSMAPLR and CSMAPLR with VTLN prior were the two systems compared in these experiment. The experimental set-up is the same as that of Dines et al. [2010]. The baseline system is the system 'd' in Table IX of this research which has 13% word error rate (WER). This system uses 13 dimensional MCEP features with CSMAPLR adaptation using a phonetic decision tree based on minimum description length (MDL) criterion. The evaluations were carried out using Spoke 4 (S4) task of the November 1993 CSR evaluations (same as the ones used in the baseline system mentioned above). The adaptation was carried out off-line using the rapid enrollment data (for condition C3) which comprises 40 adaptation utterances

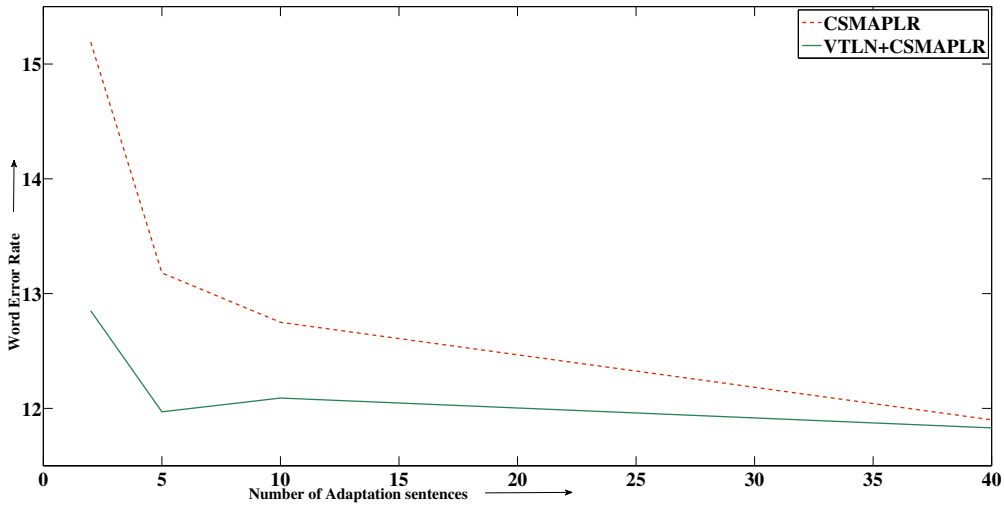


Figure 6.6 – Word error rate for CSMAPLR and the proposed VTLN-CSMAPLR.

for each of the 4 speakers. The system uses 5k word list with a bigram language model. The baseline system reported in [Dines et al., 2010] uses the value of σ, the weight of the prior as one. Increasing this value to 1000, improves the CSMAPLR WER up to 12%. The WER using different amounts of adaptation data ranging from 2 to 40 adaptation sentences are shown in Figure 6.6. It can be observed that the VTLN as prior has a very good effect on the performance when the adaptation data is as low as 2 sentences. As more adaptation data is used, the prior is no longer influential.

Experiments were performed to check the different schemes to combine VTLN with model transformations. These include the use of VTLN as a stacked transformation with CSMAPLR or using VTLN as a non-structural prior for constrained MAPLR adaptation (ignoring the tree structure). In the first case, a VTLN transformation was estimated for the given adaptation data and then, this transformation was used as a parent transformation to transform the features which were used to estimate the CSMAPLR based model transformation. In the second case, the same global VTLN transformation was used as a prior to every node of the regression tree of the CSMAPLR transformation ignoring the structural prior and thus, resulting in a constrained non-structural MAPLR (represented by CMAPLR) transformation with VTLN prior. All the systems use a prior scaling of 1000. The word error rates for these techniques for the same setup as above is plotted in Figure 6.7. The results show that the performance of the VTLN as a prior to the root of the CSMAPLR transformation still performs better than using a stacked VTLN and CSMAPLR transformation or using VTLN as a non-structural prior. The structural prior aspect of the CSMAPLR adaptation thus can be concluded to contribute a significant improvement in the ASR performance.

ASR performance is also presented for the bias parameter of VTLN. The bias parameter estimation is independent from that of the warping factor for VTLN An efficient method should be devised to combine these two components in an effective manner. For these experiments

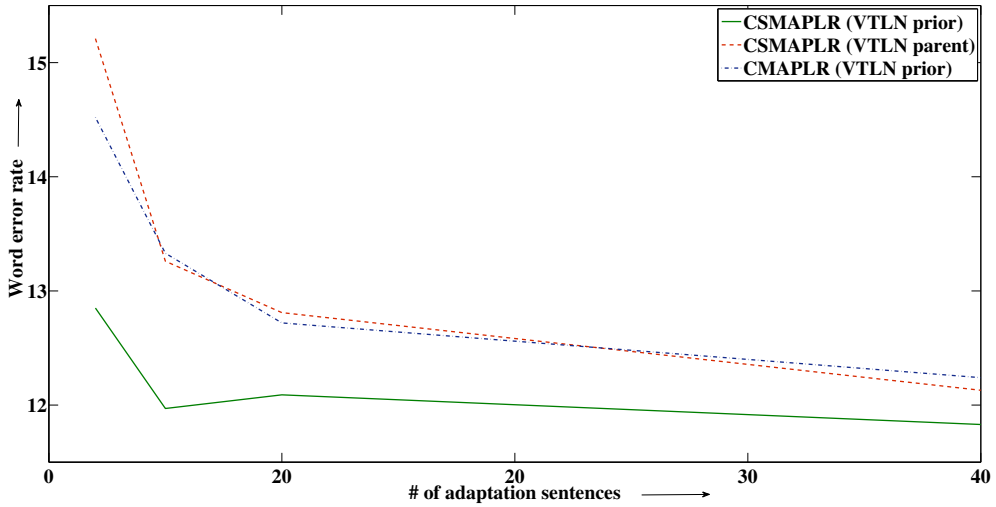


Figure 6.7 – Word error rate for VTLN stacked with CSMAPLR and the VTLN-CMAPLR systems.

VTLN warping factor ($[\boldsymbol{A}_\alpha,0]$) was estimated initially and then this transformation was used as a parent transform to estimate the bias term ($[\boldsymbol{I},\boldsymbol{b}]$ where $\boldsymbol{I}$ is the identity matrix). Finally, the two components were combined into a single transformation matrix ($\boldsymbol{A}_\alpha,\boldsymbol{b}$]). Also, a bias alone transformation was estimated using [$\boldsymbol{I}$,0] as the parent transformation. ASR evaluations were performed using all the schemes. The results are plotted in Figure 6.8. The figure shows the performance of speaker independent (SI) model without any adaptation denoted as SI-model, VTLN adaptation ($[\boldsymbol{A}_\alpha.\ 0]$) using single and multiple transforms (based on regression trees) denoted as VTLN-single and VTLN-multiple, bias alone transformation ($[\boldsymbol{I},0]$) denoted as Bias and VTLN with bias ($[\boldsymbol{A}_\alpha,\boldsymbol{b}]$) using the method mentioned above denoted as VTLN-Bias The results show that bias is an important term in the VTLN transform estimation and needs further investigation. Just with the bias factor, the recognition performance improves a lot. The overall results are not as good as the previous ones because the number of parameters in the transformations (VTLN or bias) are limited compared to the CSMAPLR transformations.

6.5 Summary of Contributions

It can be concluded that the VTLN prior can significantly improve the CSMAPLR adaptation performance when the adaptation data is very limited and unlike VTLN, can scale up to the performance of CSMAPLR with more adaptation data. This chapter has presented a few novel ideas for combining the merits of CSMAPLR and VTLN adaptation, resulting in an improved adaptation technique. An efficient algorithm was presented to use the VTLN transformation matrix as prior information for the existing CSMAPLR adaptation. Performance improvements were shown, especially when very little adaptation data was available. Speech synthesized using this technique is not only better in naturalness, but has improved intelligibility and overall better quality. A few of the techniques presented in this chapter were published as Saheer et al. [2012c].

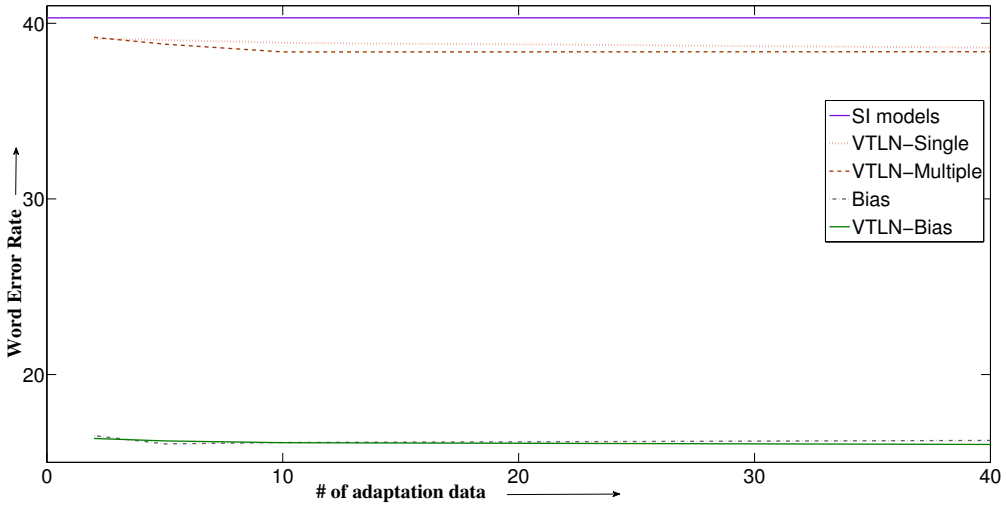


Figure 6.8 – Word error rate for bias term of VTLN.

7 Evaluating VTLN in special scenarios

The details of VTLN as a rapid adaptation technique was presented in the earlier chapters. It was seen that VTLN generates speech with better naturalness with very little adaptation data and gives additional improvements when combined with other linear transformations. It should be noted that VTLN is a very constrained transformation with a few parameters to represent the speaker variability. This constraint, though a hindrance in improving the performance to some extend, could prove to be advantageous in some special scenarios. There might be a large variability to be represented in these cases that some constraints (rather than using a large number of parameters) could result in a better transformation. This chapter explores some specific scenarios where VTLN is a lot more effective in improving performance. These situations include cross-gender transformations, child speech or age transformations, cross-lingual transformations, and noisy speech or environmental transformations. This chapter presents speech recognition or synthesis performance in these different conditions.

7.1 Related Work

There are a number of references in the literature that show performance enhancement using VTLN in ASR. Most of these publications perform evaluations on special databases like child speech, cross-gender evaluation, or noisy speech. The most commonly used database is the "TIDIGITS" (a speaker-independent connected-digit database) with a number of speakers in different age groups and categories (man, woman, boy, and girl). A number of studies like that of Panchapagesan and Alwan [2009], Sanand et al. [2009], Akhil et al. [2008], Rath et al. [2009], Miguel et al. [2006, 2008] or Karhila et al. [2012] use this database with cross-gender and cross-age evaluations like child speech (boy or girl) used as test data on the models built using adult (male or female) speech data. The results are promising when the mismatch is higher. There are evaluations with noisy data, like the "Aurora" database or "Speech-Dat-Car" database or the "OGINumbers" (multispeaker telephonic database). The work of Faria and Gelbart [2005], Miguel et al. [2005], Rose et al. [2006], Breslin et al. [2010] or Sanand and Umesh [2012] in this area use these noisy speech databases to demonstrate the power of VTLN in the environmental mismatch conditions. The list presented here is just a representative one and

there are other works in the literature that use same or similar special databases to evaluate VTLN.

7.1.1 Cross-lingual speaker adaptation

The ability to transform voice identity in TTS has been an important area of research with applications in medical, security and entertainment industries. One specific application that has seen considerable interest by the speech research community is that of speech-to-speech translation, where the challenge of voice transformation is further compounded by the differing languages of target speaker data and output synthesis. A range of speaker adaptation techniques for voice transformation in a cross-lingual setting is commonly referred to as cross-lingual speaker adaptation (CLSA).

CLSA takes speech data in one language and uses this to adapt a set of acoustic models for synthesis in a different language. Unlike in intra-lingual speaker adaptation, it is evident that the correspondence between adaptation data and the acoustic models to be adapted is largely lost at the linguistic level. To date, the most successful approaches have relied on the construction of a set of mapping rules between acoustic model distributions (i.e. HMM states) for the two languages, thus establishing sub-phonemic (or senone-level) correspondence between the two languages Wu et al. [2009]. Given this state mapping, CLSA may be performed using conventional speaker adaptation techniques such as CSMAPLR.

Despite the progress that has been made in CLSA, it is evident that the state of the art still lags behind intra-lingual speaker adaptation in terms of synthesis performance (i.e. speaker similarity, speech naturalness, or intelligibility.). This is in large part due to the fact that the state-level mapping is still unable to fully account for the inherent mismatch between phonetic inventories of different languages Liang and Dines [2010]. In order to avoid the severe effects of such a mismatch, it may be beneficial to further constrain adaptation transforms. VTLN provides one such mechanism for constraining model adaptation, where typically only a single parameter is estimated. Furthermore, vocal tract length may be considered to be inherently language independent, hence, VTLN may not suffer from such mismatch issues.

CLSA remains a challenging task; relevant literature is sparse as the field draws on several disparate concepts, each non-trivial in its own right [Oura et al., 2010, Gibson and Byrne, 2010, Liang et al., 2010]. Previous work on CLSA normally employed CSMAPLR or related adaptation techniques. In the context of intra-lingual speaker adaptation, CSMAPLR has proven effective in capturing main speaker characteristics, but its application in a cross-lingual context has met with less success, especially when multiple adaptation transforms are used [Liang and Dines, 2010]. By contrast, VTLN has significantly fewer parameters (typically only one parameter is used to modify the vocal tract warping function) and as such the range of speaker characteristics that can be represented is restricted. However, in the cross-lingual scenario, where CSMAPLR is susceptible to learning not only speaker characteristics, but also undesirable language mismatches, VTLN may provide more acceptable results. The

fact that CSMAPLR and the current VTLN implementation operate on the underlying HMM distributions in the same manner (i.e. as maximum *a posteriori*/likelihood linear feature transformation) provides a good basis for testing this hypothesis. The following sections will give the details of the evaluations performed using VTLN for the special scenarios mentioned earlier including the cross-lingual transformations and validate the hypothesis that VTLN is useful as a rapid adaptation technique.

7.2 Cross-lingual Transforms

It can be seen from earlier chapters that VTLN-synthesis was able to produce naturalness ratings close to the average voice and significantly better than model adaptation techniques like CSMAPLR while still improving speaker similarity over the average voice. In this section, a new framework facilitating supervised rapid CLSA is presented, where HMM state mapping is integrated into bilinear transform-based VTLN. The hypothesis that the constrained nature of VTLN transformation might help to alleviate some problems associated with current CLSA approaches is tested. Experiments were performed on the Mandarin-English language pair and VTLN adaptation is compared with CSMAPLR.

7.2.1 Integration of State Mapping-Based CLSA into VTLN

HMM state mapping, which has proven effective for CLSA as shown by Wu et al. [2009], was integrated into bilinear transform-based VTLN. First of all, the language in which speech is synthesized is defined as the *output language* and the language of given adaptation utterances from a target speaker is defined as the *input language*. Two monolingual average voice model sets were established in the input and output languages respectively, $\mathbb{S}^{\text{in}} = \{S_1^{\text{in}}, S_2^{\text{in}}, \cdots, S_{N^{\text{in}}}^{\text{in}}\}$ and $\mathbb{S}^{\text{out}} = \{S_1^{\text{out}}, S_2^{\text{out}}, \cdots, S_{N^{\text{out}}}^{\text{out}}\}$, where S refers to state distributions. Following this, a set of state mapping rules, $\mathbb{M}(\cdot)$, were constructed such that

$$\mathbb{M}\left(S_i^{\text{in}}\right) = \underset{S_j^{\text{out}} \in \mathbb{S}^{\text{out}}}{\arg\min} D_{\text{K-L}}\left(S_i^{\text{in}}, S_j^{\text{out}}\right), \ \forall S_i^{\text{in}} \in \mathbb{S}^{\text{in}} \tag{7.1}$$

where $D_{\text{K-L}}(\cdot,\cdot)$ denotes the symmetric Kullback-Leibler divergence between two Gaussian distributions. The state distributions comprise single Gaussian PDFs (as is usual for HMM synthesis).

Wu et al. [2009] proposed two ways of applying these state mapping rules: data transfer and transform transfer. It has been observed by Liang et al. [2010], Liang and Dines [2010] that data transfer is preferred over transform transfer, thus, this work is based on data transfer and a cross-lingual warping factor $\hat{\alpha}_s$ is estimated as follows, in a similar fashion to Eq. (2.2) (the

intra-lingual version):

$$\hat{\alpha}_s = \arg\max_{\alpha} p\left(\boldsymbol{A}_{\alpha_s}\boldsymbol{x}_s^{\text{in}} \mid \boldsymbol{\mu}^{\text{out}}, \boldsymbol{\Sigma}^{\text{out}}, \mathbb{M}\left(\boldsymbol{q}^{\text{in}}\right)\right) \tag{7.2}$$

where $\boldsymbol{x}_s^{\text{in}}$ is acoustic feature vectors of adaptation data of speaker s, $\boldsymbol{q}^{\text{in}}$ consisting of $\{S_i^{\text{in}}\}$ is the state sequence of $\boldsymbol{x}_s^{\text{in}}$, $\boldsymbol{\mu}^{\text{out}}$ and $\boldsymbol{\Sigma}^{\text{out}}$ are mean vectors and covariance matrices of an average voice in the output language.

Using more transforms is generally beneficial to the performance of intra-lingual speaker adaptation. Interestingly, Liang and Dines [2010] discovered that the fact was just the opposite in CLSA: It was better to estimate only a single global transform for all state emission distributions when using data transfer. This work also investigates whether this phenomenon will be observed in VTLN-based CLSA.

7.2.2 Investigation

The experiments performed in this section are mainly focussed on testing two hypotheses:

1. As a highly constrained feature transformation, VTLN can perform better than CSMAPLR in a rapid CLSA scenario where limited adaptation data is available.
2. Multiple transform-based VTLN also degrades performance in the cross-lingual scenario, as has been previously observed for CSMAPLR.

In this work the Mandarin-English language pair was used, with Mandarin/English being the input/output language. One Mandarin adaptation utterance and its context-dependent labels were used to generate speaker-specific transforms. The techniques compared were global/multiple VTLN transform and global/multiple CSMAPLR transform based adaptation. A global VTLN transform corresponded to a single speaker-specific warping factor applied to an entire model set. Multiple VTLN transforms corresponded to different speaker-specific and phoneme class-dependent warping factors generated from a regression class tree in the usual fashion. Likewise, a global CSMAPLR transform applied to an entire model set and multiple CSMAPLR transforms were regression class-dependent. The prior weighting for the CSMAPLR transforms were adjusted to an empirically determined value [1] of 1000, which has been previously observed (in chapter 6) to give the best results with a small amount of adaptation data [Miyamoto et al., 2009].

Experimental Setup

Two average voice synthesis models were trained on the Speecon (Mandarin, 12.3 hrs) and WSJ0 (SI-84, American English, 15.0 hrs) corpora in the HTS-2007 framework [Yamagishi et al., 2009b]. The HMM topology was five-state (single mixture component, multivariate Gaussians) and left-to-right with no skips. Speech features were 39th-order mel-cepstra, log

1. The HTS variable `HADAPT:SMAPSIGMA` was set to 1000.

F0, five-dimensional band aperiodicity, and their delta and delta-delta coefficients, extracted from 16kHz recordings with a window shift of 5ms. The setup of the following experiments is the same as that of Liang et al. [2010], except for the source of adaptation and evaluation data.

Detailed evaluations were performed on a pilot corpus recorded in an anechoic studio in the University of Edinburgh by a male, native Mandarin speaker uttering Mandarin and reasonably natural-sounding English. Only one Mandarin adaptation utterance of 7.71 seconds was used for transform estimation in all cases. In addition, a limited number of systems were selected for further evaluations with one male and three female speakers from the EMIME bilingual (Mandarin-English) corpus from Wester and Liang [2011a] recorded in the same anechoic studio. These four speakers were the ones with the least accented spoken English amongst all the speakers in this EMIME bilingual corpus; only a single Mandarin adaptation utterance of similar duration was used for each of them.

This work focuses on the cross-lingual adaptation of the spectrum. The subjective evaluations were based on ABX and AB tests for speaker similarity and naturalness, respectively. Listeners were presented with two speech samples at a time and asked to judge which one sounded closer to the voice of a reference sample (ABX test) or more natural (AB test). Mandarin reference samples were presented to the listeners for judging speaker similarity of synthesized speech in English in the ABX test. The listening tests were performed only on selected pairs of systems that could give the most useful insights with respect to the hypotheses presented earlier. Four pairs of systems are compared to each other:

1. Global transform based VTLN versus average voice speech.
2. Global transform based VTLN versus multiple transform based VTLN.
3. Global transform based VTLN versus global transform based CSMAPLR.
4. Multiple transform based VTLN versus multiple transform based CSMAPLR.

The first combination should demonstrate if the VTLN can bring in any speaker specific characteristics. The second combination should demonstrate if the multiple transform will degrade the performance as with CSMAPLR. The third combination compares the performance improvements with (constraints of) VTLN over CSMAPLR. The fourth combination shows if the multiple VTLN transforms still be able to preserve the performance improvements.

7.2.3 Evaluation Results and Discussions

It is hypothesized that VTLN produces far more natural-sounding speech than CSMAPLR, since the adaptation of a single parameter prevents gross modification of the average voice model, thereby maintaining the better naturalness of the original average voice model as shown by Yamagishi et al. [2010]. VTLN transforms spectrum only. All the other kinds of features were transformed using transforms obtained from CSMAPLR.

The initial evaluations were conducted with 4 pairs of systems for the male speaker from the pilot bilingual corpus. 37 and 28 listeners participated in the AB and ABX tests respectively.

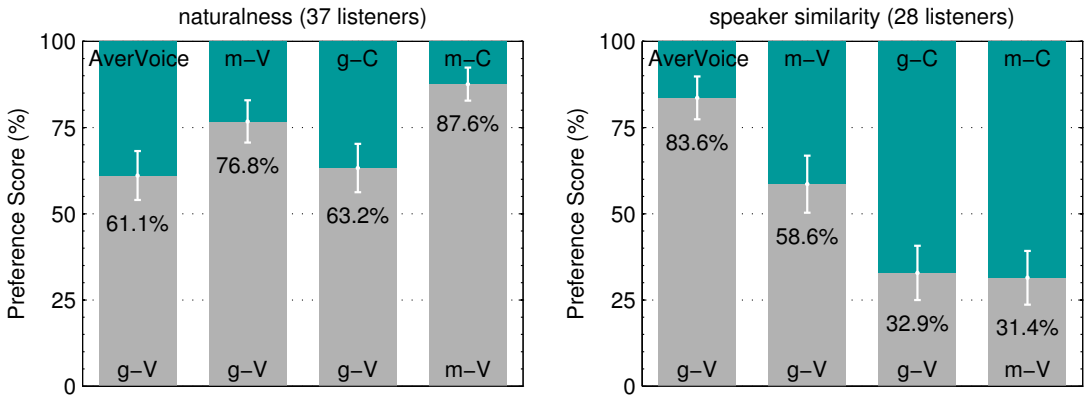
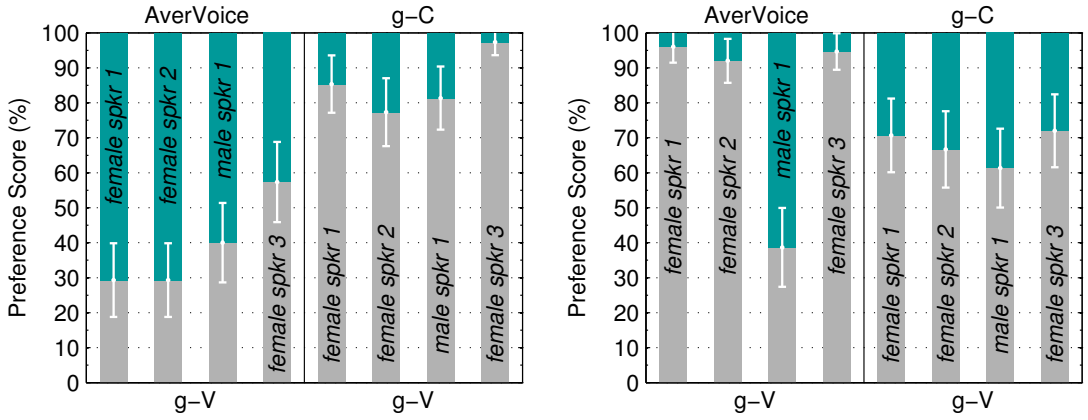


Figure 7.1 – Results for the pilot male speaker from the EMIME bilingual corpus. The systems are named as (g/m)-(V/C): g/m means *global/multiple*, and V/C means VTLN/CSMAPLR. Whiskers indicate 95% confidence intervals.

Each listener evaluated 80 English utterances in total.The results are plotted in Figure 7.1. It is evident from these figures that VTLN is far more natural compared to CSMAPLR, but the ability to achieve good speaker similarity with VTLN alone is limited. Using multiple transforms in VTLN is not as degrading as in CSMAPLR.

Based on this result and other observations, it can be postulated that the effectiveness of VTLN as a speaker adaptation technique for TTS is dependent on the characteristics of a target speaker – some speakers cannot be sufficiently reproduced using VTLN adaptation while others can. To that end, evaluations were performed with the four speakers from the EMIME bilingual corpus. Only two pairs of systems (average voice vs global-VTLN and global-VTLN vs global-CSMAPLR) were compared for these speakers for finding the effectiveness of VTLN as an adaptation technique. Each listener was presented with 20 pairs of sentences for each of the four speakers, judging naturalness and speaker similarity. Results are plotted in Figure 7.2.

Similar trends are observed in these results. Since the training data for estimation of average voice is dominated by male speakers, good results are observed with VTLN (that captures the gender characteristics) for female test speakers. VTLN cannot capture any speaker characteristics other than gender and hence, the characteristics captured for a male test speaker with this model may not be very obvious. Unlike the previous case, the VTLN system is preferred over CSMAPLR, even for speaker similarity, mainly because VTLN-synthesized speech sounded more natural than CSMAPLR. To further elaborate, neither adaptation technique could exactly reproduce a target speaker's voice characteristics with so little adaptation data. Many listeners may have unconsciously considered naturalness as a key factor of speaker similarity since the source of the reference speech is an original recording of the target speaker. Hence, the listeners preferred more natural-sounding speech. For the same reason, some male speakers could be judged closer to the average voice in speaker similarity since the average voice is male dominant and better in naturalness when compared to VTLN.



a) naturalness (15 listeners)

b) speaker similarity (15 listeners)

Figure 7.2 – Results for the four target speakers from the EMIME bilingual corpus. The systems are named as (g/m)-(V/C): g/m means *global/multiple*, and V/C means VTLN/CSMAPLR. Whiskers indicate 95% confidence intervals.

It should be noted that VTLN doesn't include any linguistic information in its transforms and thus cannot perform well for speakers with speaking styles very different from the average voice model. Nonetheless, VTLN is still useful: since it can provide natural-sounding speech close to a target speaker's voice in a language, even using a *single* adaptation utterance in another language from the speaker. It can be expected that listeners will prefer to listen to VTLN-adapted speech when very little adaptation data is available since VTLN brings in at least the gender characteristics with better naturalness. It is also worth noting the results of perception experiments of Wester and Liang [2011b], which suggest that the correctness of speaker discrimination is only 51-61% if two speech samples for comparison are in different languages *and* of different speech types (i.e. natural or speaker-adapted). Thus, judgement of speaker similarity in a CLSA context is already a difficult task regardless of the approach employed. By contrast, the advantages offered by VTLN-based CLSA with respect to naturalness are quite clear, while the approach still maintains gross speaker qualities (e.g. gender or age.). The results thus confirm the first hypothesis made in the previous section that VTLN performs better compared to CSMAPLR in a rapid CLSA scenario.

Concerning a comparison of global and multiple transform adaptation approaches, it is clear from the subjective evaluation that multiple transforms provide inferior CLSA performance. This is consistent with earlier studies of Saheer et al. [2010a], Liang et al. [2010] which showed that while multiple transforms improve the performance of intra-lingual speaker adaptation, a degradation in CLSA performance is observed. It is also noted that, based on subjective evaluation, multiple transform VTLN-based CLSA was preferable in comparison to multiple transform CSMAPLR. The second hypothesis presented at the beginning of this section also proves to be correct.

7.3 Gender Transforms

Another mismatch scenario where VTLN can perform well is the wide variation in the speakers used in training and adaptation. This variation might be because speakers are from different genders for training and adaptation. The hypothesis is that VTLN can represent the differences in vocal tract length across gender and prove to be beneficial in such scenarios. Impact of VTLN prior on CSMAPLR transformation was tested on two different databases.

i) Gender dependent models of WSJ0 American English speech recognition database.

ii) Gender dependent models of a high quality speech synthesis database.

The WSJ0 is an American English database designed for speech recognition with a large number of male and female speakers. The data contains speech waveforms sampled at 16kHz and were used to build synthesis models with 39 dimensional cepstral features. Independent male and female average voice models were generated with 59 male speakers and 60 female speakers of the WSJ0 database. A male and female test speaker was evaluated on both these models to check the influence of VTLN prior for CSMAPLR adaptation.

A database was recorded at the Centre for Speech Technology Research (CSTR), Edinburgh at the 96kHz sampling rate in the specialized anechoic recording rooms. This TTS database consists of native UK English speakers (with 31 male and 29 female speakers) and was used to build the gender dependent male and female models. The speakers were down-sampled to 48kHz and 59 dimensional cepstral features were used for the models. The evaluations were performed on 61 UK English test speakers (31 males and 30 females).

7.3.1 Results and Discussion

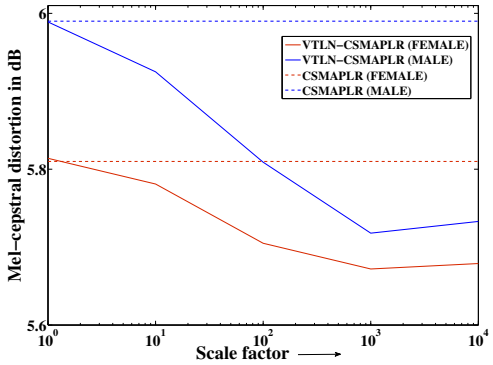
Objective scores using mel-cepstral distortion (MCD) is calculated with different amounts of adaptation data.

WSJ0 database

240 sentences each for the male and female test speakers were generated using each of the gender dependent male and female models. Different values of scale factors can be used for VTLN prior while estimating the adaptation transformation. The default value for prior scaling in HTS software is one which corresponds to no scaling. Experiments were performed with different scale factors for the VTLN-CSMAPLR combination to empirically determine the best scale factor. Figure 7.3 shows the effect of different prior scales with the MCD scores for adaptation with a single sentence for male and female acoustic models. It was observed that the scale factor of value 1000 gives the best performance in this case. Objective results for the evaluations on the WSJ0 gender dependent models are presented in Figure 7.4 and Figure 7.5. The figures show MCD scores for both male and female test speakers for each of the gender dependent model. The figure shows the MCD score for two different scale factors, 1 and 1000. The values in the bracket for the labels represent the scale factor. In all cases, using

Figure 7.3 – MCD for different scale factors for gender dependent models. The abscissa is in log scale.

(a) Male model



(b) Female model

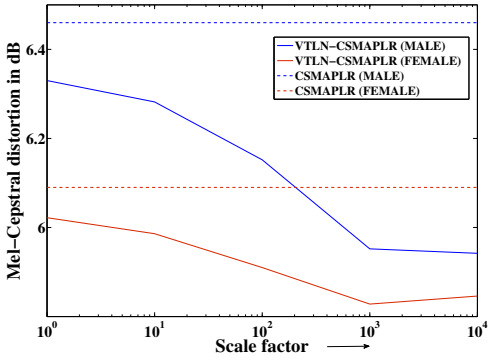
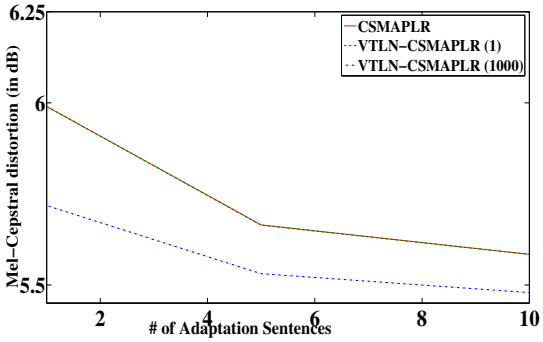
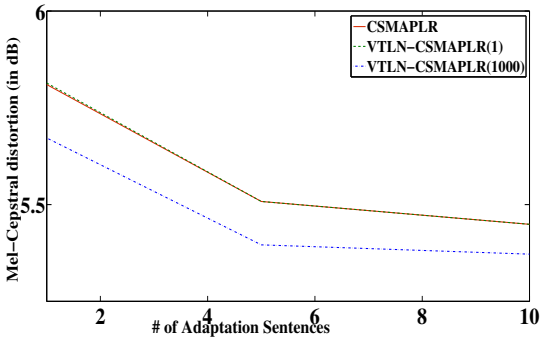


Figure 7.4 – MCD for different number of adaptation sentences for gender dependent male model

(a) Male test speaker



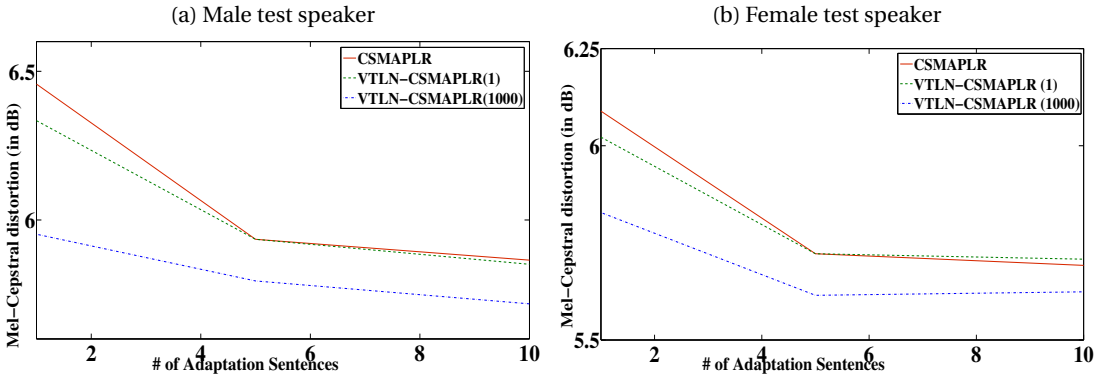
(b) Female test speaker



VTLN as CSMAPLR prior has the best performance especially when the scale factor is of the order of 1000. The results prove the hypothesis that VTLN as a gender transform gives better performance when used as a prior distribution for CSMAPLR.

Experiments were also performed to test the structural and non-structural influence of VTLN prior. The hypothesis is to check if TTS shows a performance gain from the structural prior as shown by the ASR experiments in the previous chapter. A single sentence was used as adaptation data for the male and female test speakers who were evaluated against the male and female models. As explained in the earlier chapter, VTLN as a non-structural prior involves using VTLN transforms as prior for each node of the regression tree used in the CSMAPLR transformation. Since the structural prior does not exists any more, this technique was termed as CMAPLR instead of CSMAPLR. A global VTLN transform can be used as the prior at each node or regression class based multiple VTLN transforms can be used as prior for different nodes of the tree. In the second case, different nodes will have different VTLN transforms as

Figure 7.5 – MCD for different number of adaptation sentences for gender dependent female model



prior. The set-up is exactly similar to the earlier experiment, 240 sentences were generated for each male and female test speaker. The objective measures as MCD scores are shown in Table 7.1. The table shows the VTLN as CSMAPLR prior at the root node (VTLN-CSMAPLR), global VTLN transform as non-structural prior at each node of the tree (VTLN-CMAPLR-global) and multiple VTLN transforms as non-structural prior at each node of the tree (VTLN-CMAPLR-multiple). It can be seen that the MCD scores do not vary more than 0.1dB in the different cases and hence, the systems can be judged as equivalent to each other. Unlike the results in ASR, it cannot be shown that the gain was from the structure or the VTLN prior as such. Thus, the hypothesis is wrong in the case of TTS where the structure and non-structural VTLN prior gives similar performance. The structural prior is still proposed as a better technique as it has a better prior being propagated and is shown to give performance improvements for ASR and also for other studies in the literature.

Techniques	**Male Model**		**Female Model**	
	Male Test	Female Test	Male Test	Female Test
VTLN-CSMAPLR	5.9896	5.8143	6.3339	6.0223
VTLN-CMAPLR-global	5.9883	5.7973	6.3757	6.0032
VTLN-CMAPLR-multiple	5.9869	5.7983	6.3663	6.0008

Table 7.1 – MCD for gender dependent models using structural and non-structural VTLN prior

Evaluations were performed to check the influence of bias on the rapid adaptation. Bias has to be combined with VTLN and the CSMAPLR transformations to achieve the best rapid adaptation performance. Two methods were adopted to this end. Bias can be seen as cepstral shift similar to cepstral mean normalization. Hence, the first method involves using bias as a cepstral normalization technique for the model means and then continue as before using VTLN as prior for CSMAPLR transformation. Second method involves estimating VTLN and bias iteratively with one as a parent transform of the other and finally combining them both into a single transformation. This combined VTLN and bias transformation can act as the prior for CSMAPLR. The hypothesis here is that bias term can add improvements to rapid

adaptation when combined with VTLN and CSMAPLR. Objective evaluations as MCD scores were estimated on the gender dependent female models. Both male and female test speakers were evaluated. The results compare four different systems:

1. VTLN and bias estimated iteratively with each one as a parent transform of the other. This transformation (referred to as VTLN+Bias) is used to adapt the model.
2. VTLN is used as a prior to the root node of the CSMAPLR transformation (referred to as VTLN-CSMAPLR). This is the same system presented in earlier experiments.
3. Bias used as a cepstral mean normalization and then, VTLN transformations estimated and used as prior at the root node of the CSMAPLR transformation. This system is named BiasCN-VTLN-CSMAPLR
4. VTLN along with Bias as presented in the case 1 being used as a prior at the root node of CSMAPLR transformation (referred to as VTLN+Bias-CSMAPLR)

Techniques	**Male Test**		**Female Test**	
	One sentence	five sentences	One sentence	five sentences
VTLN+Bias	6.0952	6.0332	6.0197	5.9827
VTLN-CSMAPLR	5.9525	5.6150	5.8285	5.7947
BiasCN-VTLN-CSMAPLR	6.0977	5.7830	5.8999	5.6304
VTLN+Bias-CSMAPLR	5.9795	5.7950	5.8653	5.6148

Table 7.2 – MCD for gender dependent female models using bias for VTLN prior

The results are shown in Table 7.2 for one and five adaptation sentences. The results show no perceivable difference when bias is combined using the techniques proposed. The hypothesis cannot be established. This could be because when acting just as a prior, bias term is not able to contribute to performance enhancement. The combination of bias with VTLN and CSMAPLR requires further investigation in order to utilize full potential of the bias term.

60 speaker TTS database

Different adaptation techniques discussed in chapter 6 were compared using the gender dependent models of this database. Experiments were performed separately with the gender dependent male and female models. The techniques that were compared include:

1. **CMLLR** is a transformation that does not use any prior information.
2. **CSMAPLR (No prior)** is a transformation that uses structural prior and ML estimation (no prior) at the root node.
3. **CSMAPLR (Identity prior** is a transformation that uses structural prior and identity matrix as prior for the root node.
4. **CSMAPLR (VTLN prior)** is a transformation that uses structural prior and VTLN transform as prior for the root node.
5. **CSMAPLR (VTLN Parent)** is a cascade of CSMAPLR and VTLN transformations with VTLN as the parent and CSMAPLR as the child transformation.

Figure 7.6 – MCD for gender dependent male model with different adaptation schemes

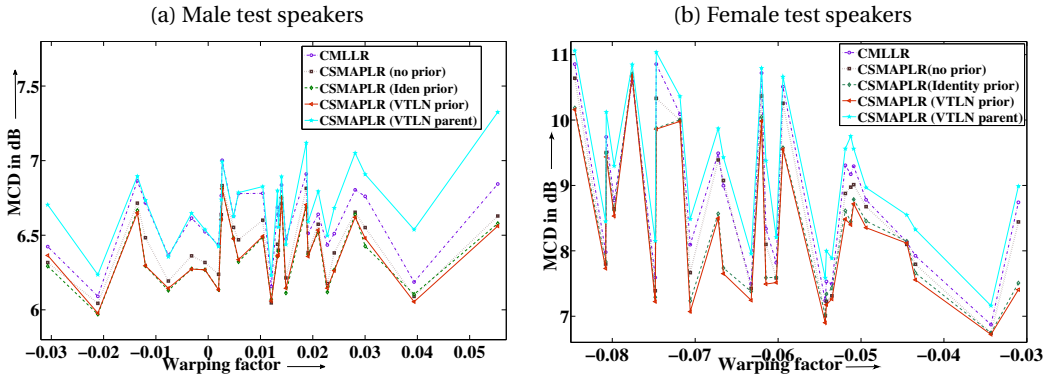
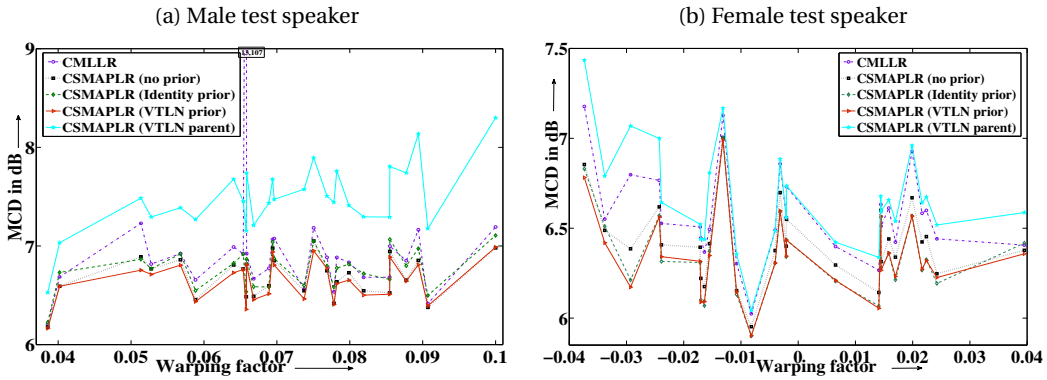


Figure 7.7 – MCD for gender dependent female model with different adaptation schemes

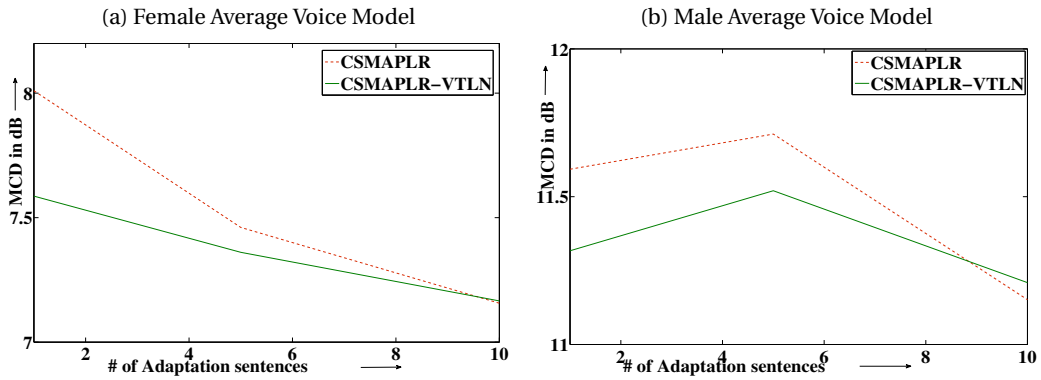


A single adaptation sentence was used to generate the transformations in each of this case. About 100 sentences were synthesized with each of these techniques for each test speaker. The CSMAPLR techniques used a prior scale of 1000 which was empirically found to be optimal. The objective results as MCD scores are plotted in Figure 7.6 and Figure 7.7. The speakers are plotted in the increasing order of their VTLN warping factor values. The figures show the performance of the male and female test speakers with each of the gender dependent male and female models. The results show that VTLN as CSMAPLR prior gives the least MCD value and overall best performance in all cases. The identity prior for CSMAPLR also performs well in all the cases. This emphasizes the fact that prior is important factor in the CSMAPLR transformations and further, an appropriate prior like a VTLN transformation can further improve performance.

7.4 Age Transforms

The ratio of the vocal tract length which is the estimate for the warping factor in the VTLN varies across different individuals. As mentioned earlier, it is longer in males and shorter

Figure 7.8 – MCD for Child speech with two gender dependent models

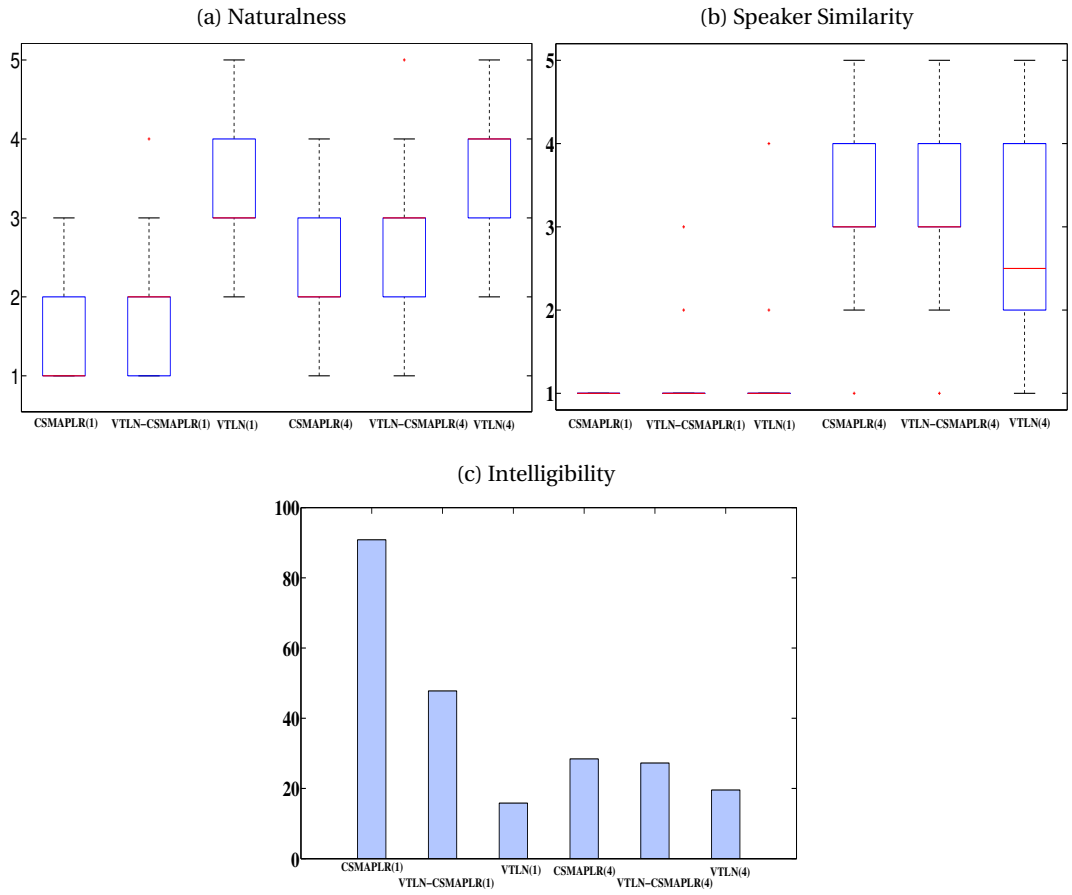


in females. This length is proportional to the actual size of the individual and hence, the vocal tract length is shortest in a child (people with smaller bodies). This poses a case of extreme or significant warping when an adult or specifically male model is adapted to a child voice for both synthesis or recognition. Children also have a higher pitch and pitch should be ideally correlated to spectral peaks. But, this study is out of scope of this thesis. Hence, only the spectral warping is taken into consideration. As mentioned earlier, most of the VTLN research seen in the literature use child speech as a main measure of showing the performance improvements with VTLN. The hypothesis is that VTLN should give performance improvements with child speech which has the greatest variation in vocal tract length. This section also presents some synthesis experiments with child speech as the adaptation data. The influence of VTLN was evaluated using both subjective and objective evaluations. The speech sampled at 48kHz were collected from two different children at the anechoic recording studio of the Centre for Speech Technology Research (CSTR), Edinburgh. The children were asked to read fairy tales. Only the data from one child was manually annotated to have a full set of reference data for objective evaluations. Other child had only four annotated sentences for adaptation. This child was used in the subjective evaluations.

The objective evaluations were performed on the gender dependent male and female models (presented in the earlier section). The 48kHz speech samples were used to extract 59 dimensional static mel-cepstral features. The 5 state HMMs were trained using the static and dynamic (Δ and Δ^2) of spectrum, logF0 and band aperiodicity features. Different amounts of adaptation data was used to synthesize 100 target sentences. The results of objective evaluations as mel-cepstral distortion (MCD) scores are plotted in the Figure 7.8. The child speech has more variations from a male speaker than a female speaker. This is reflected in the overall MCD scores. The two schemes that are compared are:

1. CSMAPLR transformation with structural prior and ML estimation (no prior) at the root node
2. CSMAPLR transformation with structural prior and the VTLN matrix as prior at the root node.

Figure 7.9 – Subjective evaluations for child speech



All systems use a prior scaling of 1000. MCD scores are presented for one, five and ten adaptation sentences. It can be seen from the results that the VTLN has very good influence when the amount of adaptation data is as little as one sentence. As the amount of adaptation data increases, both systems start to converge. Using VTLN can considerably improve the rapid adaptation performance for child speech synthesis. This is in agreement with the literature showing the influence of VTLN for child speech recognition when amount of adaptation data is limited and thus proves the hypothesis..

The speaker dependent male model (data from Blizzard 2010) is used as the base model for adapting to a child speech. The subjective evaluations for naturalness, speaker similarity and intelligibility were performed on three different systems,

1. VTLN transformation.
2. CSMAPLR transformation with structural prior and ML estimation (no prior) at the root node
3. CSMAPLR transformation with structural prior and the VTLN matrix as prior at the root node.

All systems are evaluated for adaptation with one and four sentences. 14 listeners participated in these evaluations and the results are plotted in Figure 7.9. The results for naturalness and speaker similarity are plotted as MOS scores ranging from 1 (Completely Unnatural / Sounds like a totally different person) to 5 (Completely Natural / Sounds like exactly same person). The word error rates for the text typed in by the listeners after perceiving the target speech is plotted as the result for speech intelligibility. It can be observed from the results that VTLN gives the best naturalness and intelligibility scores especially for a single sentence adaptation. In this case, CSMAPLR transformation is not intelligible at all. This gives further proof to the hypothesis that VTLN is useful as a rapid adaptation performance in child speech synthesis.

7.5 Environmental Transforms

In this section VTLN is evaluated in the context of noisy speech environment. The model transformations like CMLLR or CSMAPLR, capture some environment characteristics rather than just speaker characteristics. The hypothesis is that the constraints imposed by the VTLN transformation ignores the parameters of the environment and might be useful in noisy speech conditions. Both ASR and TTS experiments are presented in this section to validate this hypothesis.

7.5.1 Noisy Speech Recognition

The Aurora4 database represents the noisy version of the WSJ0 database. The recognition models built using WSJ American English database can be used to recognize the Aurora database. Similar to the ASR experiments presented in chapter 6, the models were in accordance with the unification theme for ASR and TTS. The 13 dimensional MGCEP coefficients were used to generate the HMMs with single component Gaussian PDFs. The experimental setup was same as that in Dines et al. [2010]. Apart from the clean condition, there are six different noise types in the Aurora4 database viz., car noise, babble noise, street noise, airport noise, restaurant noise, and train noise. Evaluations were performed using different amounts of adaptation data ranging from 2 to 40 adaptation sentences. The systems that were compared are the CSMAPLR adaptation with the CSMAPLR-VTLN which uses VTLN as the prior at the root node of the CSMAPLR adaptation. The models were build using clean speech and the adaptation data and test data were noisy representing a mismatched training condition.

The results as word error rates in different noise conditions are plotted in Figure 7.10. The evaluations are performed on different amounts of adaptation data as plotted in the figure. It can be seen from the results that VTLN as prior to CSMAPLR gives considerable improvements to all types of noise especially when the amount of adaptation data is less than ten sentences. As more and more adaptation data comes in, the prior does not have much effect and the performance of CSMAPLR is maintained. The prior scaling factor used is empirically optimized to a value 1000. It can also be noted that the overall performance of the Aurora4 database cannot be compared with the state of the art results. This is due to the fact the features (MGCEP

is not as good as MFCC for ASR) and models are aligned to the unification of TTS and ASR and not exactly the perfect setup for an ASR system. Furthermore, in order to evaluate the exact influence of VTLN, the models or the features do not use any noise reduction techniques like cepstral mean or variance normalization.

7.5.2 Noisy Speech Synthesis

As a part of the EMIME project, data was collected at a conference venue. People were asked to speak out some text in the background of a conference interaction. This resulted in the data with babble noise. Many speakers participated in this data collection. There were fewer female speakers compared to the male speakers. Forty five speakers (including six females) speaking the same text and with adequate adaptation data were selected to perform these evaluations. The models were build using the clean TTS database, the one that was collected at the specialized anechoic recording studios of the Centre for Speech Technology Research (CSTR), Edinburgh. This is the same database used in the gender dependent experiments. This TTS database consists of native UK English speakers (with 31 male and 29 female speakers) and was used to build the gender dependent male and female models. The speech recorded at 96kHz were down-sampled to 48kHz and 59 dimensional cepstral features were used for the models. Again, the adaptation and reference data are noisy and the models were build using clean speech representing the mismatched training conditions. Different systems are compared in this experiment similar to the ones in the earlier gender dependent experiments.

1. **CMLLR** is a transformation that does not use any prior information.
2. **CSMAPLR (No prior)** is a transformation that uses structural prior and ML estimation (no prior) at the root node.
3. **CSMAPLR (Identity prior** is a transformation that uses structural prior and identity matrix as prior for the root node.
4. **CSMAPLR (VTLN prior)** is a transformation that uses structural prior and VTLN transform as prior for the root node.
5. **CSMAPLR (VTLN Parent)** is a cascade of CSMAPLR and VTLN transformations with VTLN as the parent and CSMAPLR as the child transformation.

A single adaptation sentence was used to generate the transformations in each of this case. About 100 sentences were synthesized with each of these techniques for each test speaker. The CSMAPLR techniques used a prior scale of 1000 which was empirically found to be optimal. The objective results as MCD scores are plotted in Figure 7.11b and Figure 7.11a. The speakers are plotted in the increasing order of their VTLN warping factor values. The figures show the performance of the test speakers with each of the gender dependent male and female models. The results are consistent with the observations made earlier with the gender dependent experiments and show that VTLN as CSMAPLR prior gives the least MCD value and overall best performance in all cases. The identity prior for CSMAPLR also performs

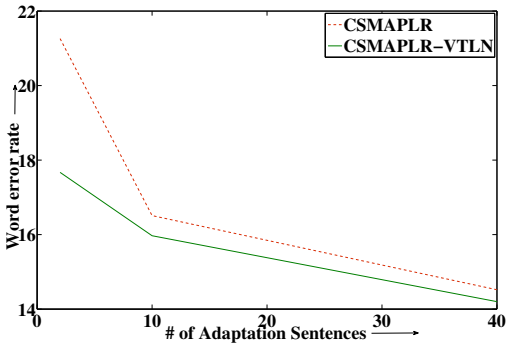
well in all the cases. This emphasizes the fact that prior is important factor in the CSMAPLR transformations and further, an appropriate prior like a VTLN transformation can further improve performance.

7.6 Summary of Contributions

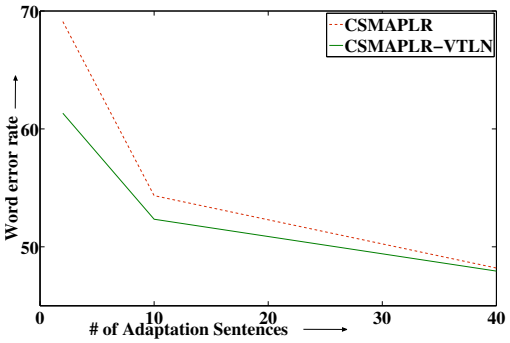
This chapter presents some special cases where the model transformation based adaptation has additional challenges of representing extreme variations between the model and the sadaptation data. These scenarios include gender transformation, transformation to child speech, transformation using noisy adaptation data, and transformation using adaptation data in another language (cross-lingual adaptation). The model transformations like CSMAPLR captures other differences in the training and adaptation data (like environment or language) rather than just speaker characteristics. VTLN is a constrained transformation and is shown to be useful in these cases. Different subjective and objective evaluations for speech synthesis and speech recognition using VTLN were performed in these special scenarios. It was shown that VTLN can add improvements to the performance especially with very little adaptation data. Combining VTLN with CSMAPLR as a prior can scale the performance to the state of the art model transformations with more adaptation data still giving considerable improvements when the amount of adaptation is scarce. The cross-lingual experiments presented in this chapter were published as a technical report [Saheer et al., 2012b] and the other experiments are planned to be submitted to a journal publication.

Figure 7.10 – WER for noisy speech database with different amounts of adaptation data

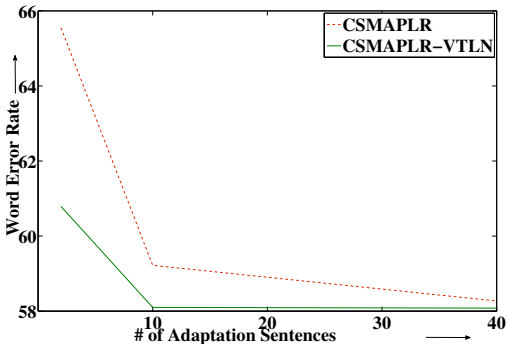
(a) Car Noise



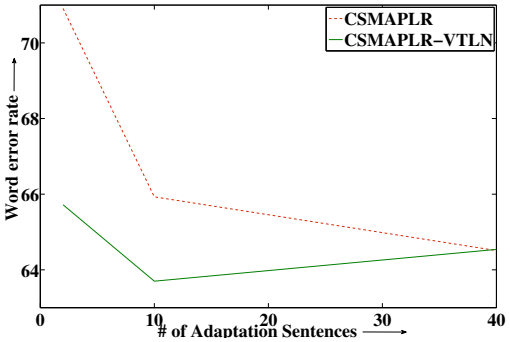
(b) Babble Noise



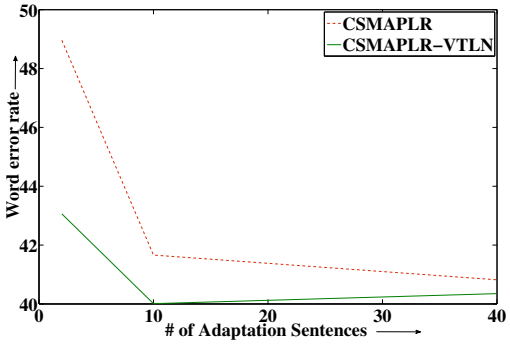
(c) Restaurant Noise



(d) Street Noise



(e) Airport Noise



(f) Train Noise

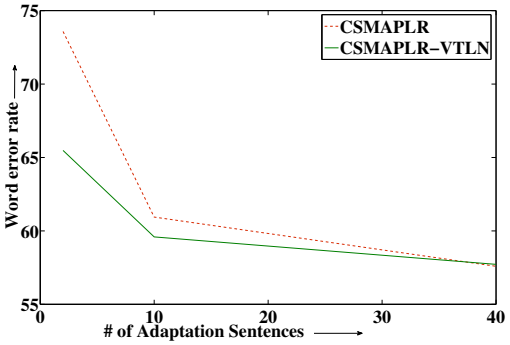
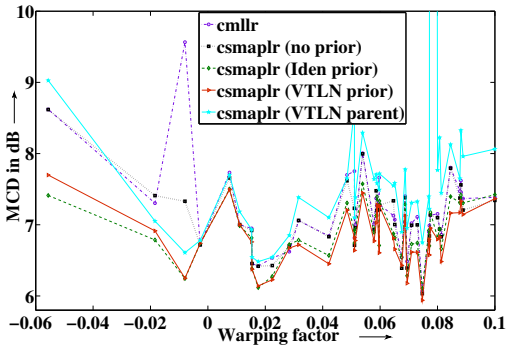
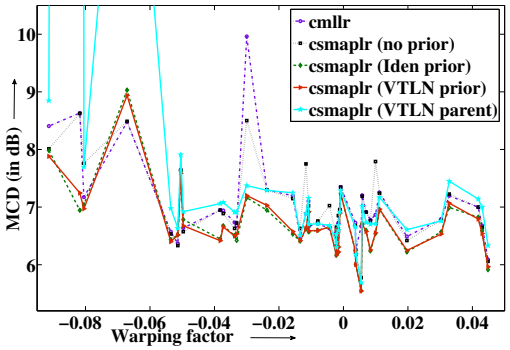


Figure 7.11 – MCD for noisy speech database with different adaptation schemes

(a) Gender dependent female model



(b) Gender dependent male model



8 Conclusions

VTLN is a well known rapid adaptation technique in ASR. Little effort has been made to date to exploit the potential of VTLN in speech synthesis. This work has explored a VTLN based rapid adaptation technique for statistical parametric speech synthesis. Recent advances in the field of HMM based statistical parametric speech synthesis have paved the way for common adaptation and modeling techniques being used for both systems. To this end, the work presented in this thesis tends to be in line with the unification theme of ASR and TTS. Though originally developed for TTS, all the techniques developed in this thesis are equally applicable to ASR where they may likewise contribute to the state of the art. Experiments were presented with both ASR and TTS (using almost the same setup - features/models) for each proposed technique in order to support the unification theme of this research.

MGCEP features are commonly used in the HMM based speech synthesis. These features include a time domain warping using all-pass transformations. An effective VTLN technique was implemented by cascading additional layers of all-pass transform based VTLN on the existing bilinear transformations of the MGCEP features. All-pass transform based VTLN has a single warping parameter that represents the ratio of an original vocal tract length to an average vocal tract length. The initial experiments were performed using a grid search warping factor estimation technique which selects the best value for the warping factors from a grid of possible values using the maximum likelihood criteria.

This implementation validated the fact that VTLN is a technique that can give performance enhancement for statistical parametric speech synthesis. VTLN was also observed to give improvements in performance when used during training of the speaker adaptive model. The main challenge with the grid search is the time and space complexity in estimating features using every warping factor in the grid. An efficient EM based warping factor estimation technique was formulated in order to overcome the complexities and to embed the warping factor estimation into the HMM training. Brent's technique was used to search for the optimal value of the warping factor using the auxiliary function of the EM method. This formulation also helped in estimating multiple warping factors for different phoneme classes using the regression class based transform estimation similar to CMLLR. Experiments were performed

on standard ASR databases like WSJ0 (American English) and WSJCAM0 (British English) to demonstrate the performance improvements with VTLN. It was shown that VTLN is a useful rapid speaker adaptation technique and multiple transformations did not give much improvements in the performance when compared to the single transformation case.

It was observed that there are some differences in the theory and practise of the VTLN even for ASR. These differences become prominent with the higher order features used in synthesis. These discrepancies include the choice of using a Jacobian normalization and problems with accurate representation of the spectrum in the case of VTLN for ASR. These differences are further compounded in statistical parametric speech synthesis with the use of higher order features and the challenges in perceptual evaluations of VTLN. A number of techniques are proposed in this work to get around this mismatch. The same techniques are expected to work equally well for ASR. The techniques include the use of a prior distribution for the warping factors resulting a MAP optimization rather than ML. Another technique was using a scale factor for the likelihood. A more promising solution is to use the lower order features (around first 13 coefficients of MGCEP features) that represent the spectral envelope to optimize the warping factors. The perceptual mismatch should be handled to some extend by choosing the target speakers for evaluating VTLN performance. It was observed that speakers with fewer differences (accent or speaking style) from the average voice model will have an advantage over using VTLN.

VTLN has a single parameter that represents the vocal tract length which is the gender characteristics of a speaker and has limited potential for transforming any other characteristics of a speaker. There are more powerful model based transformation techniques like CMLLR and CSMAPLR that have a number of parameters which could accurately represent the speaker characteristics when an adequate amount of adaptation data is available. Especially if the target speaker has great differences in accent or speaking style from the average voice models, then VTLN cannot fully represent the target speaker and also cannot scale to the performance of the model based adaptation technique with the availability of large amounts of adaptation data. In order to overcome this difficulty, it is ideal to combine VTLN with techniques like CMLLR and CSMAPLR so that there is satisfactory performance with limited amounts of adaptation data and once more adaptation data is available, the performance is similar to that of the model based adaptation techniques. Different techniques are proposed to combine VTLN with CSMAPLR based adaptation technique. These include use of VTLN as the prior at the root node of the CSMAPLR adaptation or as a prior to every node of the tree thus giving a non-structural framework or using VTLN as a parent transformation for CSMAPLR. The best technique that was shown to give the desired or even improved performance was the use of VTLN as the prior at the root node of the CSMAPLR transformation.

VTLN represents the scaling of the spectrum and misses a major component that is the translation to represent the exact spectral energies. This component is referred to as the bias in a transformation and is an important factor for the performance of an adaptation technique. The bias component of VTLN was derived so as to make up for this inadequacy. Experiments

were performed to show that bias plays a major role in the VTLN transformation. It is always tricky to estimate the bias factor since it is estimated independently of the VTLN warping factor and could give inconsistencies with the likelihood maximization. A detailed study in this area must be performed in order to harness the full potential of the bias factor.

VTLN is a very constrained transformation and this can be a boon when the targeted transformation requires some constraints like cross-lingual adaptation or a noisy speech adaptation. The model transformations in these situations might also capture the environment or language specific characteristics rather than just speaker characteristics. When the amount of adaptation data is very limited, VTLN will be able to perform better since it captures only the vocal tract length of the speaker. These hypotheses were validated by the experiments for both ASR and TTS, again, in accordance with the unification theme of the thesis. VTLN as such will perform better when the speakers are inherently diverse in gender like male or female speakers. When a gender dependent male model is used for synthesizing or recognizing a target female speaker, the VTLN transformation will prove to be beneficial. Even in the case of a child speaker who has a very small vocal tract length compared to adults will perform better with VTLN. These specific scenarios where VTLN proves to be useful are demonstrated using a few experiments.

8.1 Summary of Contributions

This research represents the first successful attempt to perform vocal tract length normalization based rapid speaker adaptation for statistical parametric speech synthesis. This work contributes a number of novel techniques and insights for speaker adaptation when the amount of adaptation data is limited. Techniques and studies presented here are equally applicable for ASR and TTS and contribute to unification theme for these two systems. The contributions include:

– A novel rapid adaptation based on VTLN is proposed using the all-pass transformations. Though originally proposed for MGCEP features, the system can be easily adopted to any feature representation that has a time domain or cepstral domain transformation with unsmoothed (like not using any filter bank smoothing) representation of the spectrum.
– The all-pass transform based VTLN using the maximum likelihood optimization criterion was initially implemented using a grid search approach.
– Once VTLN was shown to be yielding performance enhancements, an efficient EM based warping factor estimation algorithm was implemented in order to estimate accurate warping factors and improve the time and space complexities of the warping factor estimation. This implementation helped to represent VTLN as a model transformation similar to CMLLR/C-SMAPLR, even use the same sufficient statistics and embed the warping factor estimation into the HMM training. Brent's search based optimization technique is proposed using the EM auxiliary function.
– This work also proposes several successful techniques to unify the theory and practise of VTLN in both ASR and TTS. The techniques l ike using a MAP optimization of VTLN

parameters and use of lower order feature parameters for warping factor estimation prove useful in narrowing the divergence of theory and practise.
– Another set of useful and novel techniques proposed in this research are the different methods of combining VTLN with powerful model based adaptation schemes like CMLLR/CSMAPLR. VTLN is proposed to work as a parent transform for CSMAPLR or as a prior in a structural or non-structural framework for CSMAPLR. According to this study, the most successful method was the use of VTLN as the prior at the root node of the structural CSMAPLR transformation.
– This work also proposes improvements to the VTLN transformation per se using regression class based multiple VTLN warping parameters.
– Another major contribution of this work is the derivation of a bias term for VTLN transformation and using it with the existing warping factors for further enhancement of the VTLN performance.
– Finally, this study also proposes various scenarios where the contribution of VTLN can be more prominent like the cross-gender or the child (age) or the cross-lingual or the noisy speech transformations. The experiments are presented to show that VTLN proves to be an useful rapid adaptation technique especially in these special conditions.

Most of the findings in this thesis have received considerable approval and encouragement from the research community a nd are published in reputed conferences and peer reviewed journals. The different implementations of VTLN and combinations with CSMAPLR has been published as extensions to the open source HTS software.

8.2 Future Directions

There are several future directions to this work. All-pass transformations were implemented for MGCEP features as these are the most commonly used features in statistical parametric speech synthesis. The schemes described in this work should work with other feature representations that have a cepstral representation that is a linear transformation of the spectrum (i.e. not smoothed using a bank of filters). It might be interesting to try this system for such a feature representation that is optimized for ASR in order to achieve the state of the art performance. There can be other complicated all-pass transformation functions like the rational all-pass transforms (RAPT) sine-log all-pass transforms (SLAPT) which might yield better performance for VTLN as a speaker transformation as mentioned by McDonough [2000]. The algorithms mentioned in this thesis can be scaled to use these transformation functions. As such the VTLN performance can be improved by using a bias term as proposed in this work. A more detailed evaluation needs to be performed on how to effectively combine the bias term which is estimated independently from the warping factor term or even how to effectively interleave the warping factor estimation with the bias term estimation. Also, only a global bias is evaluated in this work, multiple bias terms with multiple VTLN transformations can also be estimated and tested for performance. The multiple transform based VTLN is not explored to its full potential as in using multiple priors for the different nodes of the regression class

based model transformation. More efficient and effective ways to combine VTLN with model transformations can be investigated like using a tree structure adaptation with interleaving VTLN or CSMAPLR transformations for different nodes depending on the availability of the adaptation data for each node. Even the VTLN transformation can be represented in a hierarchical structural transformation using a MAP optimization similar to CSMAPLR with a proper formulation of the prior distribution like a beta distribution. A count smoothing framework can be used for combining the VTLN prior with the CSMAPLR adaptation thus testing VTLN as a count smoothing prior for model adaptation in a structural framework. Warping factor represented by the VTLN may not be independent from the pitch contour of the speaker and might performed better when represented as in a unified formulation of the spectral and pitch transformations. Another totally diverse future direction (originally planned in this research) can be optimizing the modelling parameter, γ, of the MGCEP features in lieu of the unified modelling theme of the thesis and formulating it as an EM optimization problem along with combining it with the warping parameter (α) optimization. Maximum likelihood as an optimization criteria may not be the optimal one for VTLN on statistical parametric speech synthesis. It might be interesting to also investigate other optimization criteria, like minimum generation error, which is more inline with the synthesis theory. Some of the future directions like the optimal implementation of the bias, the structural VTLN approach, the multiple VTLN transforms as prior for CSMAPLR transformations or a count smoothing VTLN prior for CSMAPLR were not investigated in this work due to the time limitations. Some other proposed future directions like the unification of pitch, unification of the modelling factor (γ) optimization, use of other optimization criteria or other transformation functions were not investigated as they diverged from the scope of this work. In any case, all the above ideas are worth investigating for improving rapid speaker adaptation performance using VTLN.

A Deriving MGCEP Recursions

Oppenheim [1972] talks about representation of continuous signals using discrete sequence and representation of one discrete signal by another discrete representation. The main idea is that the new representation should preserve discrete convolution. The bilinear transform is one such transform that preserves convolution.

Let a continuous function be represented by two different sequences. The sequences can be represented as a combination of orthogonal complete functions. Let g_k and f_k be two different sequences representing f(t) and let $\phi_n(t)$ be complete.

$$f(t) = \sum_{n=-\infty}^{+\infty} f_n \phi_n(t) \qquad \text{and} \qquad f(t) = \sum_{k=-\infty}^{+\infty} g_k \lambda_k(t) \tag{A.1}$$

$$\lambda_k(t) = \sum_{n=-\infty}^{+\infty} \psi_{k,n} \phi_n(t) \qquad \text{and} \qquad f_n = \sum_{k=-\infty}^{+\infty} g_k \psi_{k,n} \tag{A.2}$$

The z-transform of $\psi_{k,n}$ must satisfy the relation given by Equation A.3. These equations are derived for the continuous case in Oppenheim's paper.

$$\psi_r(z)\psi_{k-r}(z) = \psi_k(z) \qquad \text{and} \qquad \psi_k(z) = [\psi_1(z)]^k \tag{A.3}$$

This implies that the z-transform of f_n and g_k can be related by substitution of variables. Let F(z) and $G(\hat{z})$ represent z-transform of f_n and g_k, then, $\hat{z} = [\psi_1(z)]^{-1} \triangleq m(z)$ and F(z) = G[m(z)].

For bilinear transform case with α as the warping factor

$$\psi_k(z) = \left(\frac{z^{-1} - a}{1 - az^{-1}}\right)^k \tag{A.4}$$

$$\hat{z} = m(z) = \frac{1 - az^{-1}}{z^{-1} - a} \tag{A.5}$$

$\psi_{k,n}$ is orthogonal with a weighting sequence 'n'. The following equation can be viewed as an expectation function.

$$\sum_{n=-\infty}^{+\infty} n\psi_{r,n}\psi_{k,n} = \begin{cases} k & \text{k = r} \\ 0 & \text{k} \neq \text{r} \end{cases}$$

$$g_k = \begin{cases} \frac{1}{k}\sum_{n=-\infty}^{+\infty} n\psi_{k.n} f_n & \text{k} > 0 \\ 0 & \text{k} < 0 \end{cases}$$

if f_n is a causal function, $f_n = 0$ for $n < 0$.

Transfer function $h_{k,n} = \frac{1}{k} n\psi_{k,n}$ and $H_k(z) = \frac{-z}{k}\frac{d\psi_k(z)}{dz}$. Substituting the equation A.4, we get,

$$H_k(z) = \begin{cases} \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}}\left(\frac{z^{-1}-\alpha}{1-\alpha z^{-1}}\right)^{k-1} & \text{k} > 0 \\ \frac{1}{1-\alpha z^{-1}} & \text{k} = 0 \end{cases}$$

This can be realized by a cascade of filters given in the paper. The bilinear transform gives an infinite length sequence which needs to be truncated.

A.1 Frequency transformation recursion from the cascade of filters

Consider an input sequence x_t and output sequence y_t with corresponding z-transforms represented by $X(z)$ and $Y(z)$. The transfer function is given by $H(z) = Y(z)/X(z) = 1/(1 - \alpha z^{-1})$ for k=0. Rearranging the variables.

$$\begin{aligned} Y(z) &= X(z) + \alpha z^{-1} Y(z) \\ y_t &= x_t + \alpha y_{t-1} \\ y^i(0) &= x^i(-i) + \alpha y^{i-1}(0) \end{aligned} \tag{A.6}$$

The transfer function is given by $H(z) = Y(z)/X(z) = (1-\alpha^2)z^{-1}/(1-\alpha z^{-1})$ for k=1. Rearranging

the variables.

$$\begin{aligned} Y(z) &= \alpha Y(z)z^{-1} + (1-\alpha^2)z^{-1}X(z) \\ y_t &= \alpha y_{t-1} + (1-\alpha^2)x_{t-1} \\ y^i(1) &= \alpha y^{i-1}(1) + (1-\alpha^2)y^{i-1}(0) \end{aligned} \tag{A.7}$$

The transfer function is given by $H(z) = Y(z)/X(z) = (z^{-1}-\alpha)/(1-\alpha z^{-1})$ for $k > 1$. Rearranging the variables.

$$\begin{aligned} Y(z) &= X(z)z^{-1} + \alpha(z^{-1}Y(z) - X(z)) \\ y_t &= x_{t-1} + \alpha(y_{t-1} - x_t) \\ y^i(m) &= y^{i-1}(m-1) + \alpha(y^{i-1}(m) - y^i(m-1)) \end{aligned} \tag{A.8}$$

The Equations A.6,A.7 and A.8 can be summarized as:

$$y^i(m) = \begin{cases} x^i(-i) + \alpha y^{i-1}(0) & \text{m} = 0 \\ \alpha y^{i-1}(1) + (1-\alpha^2)y^{i-1}(0) & \text{m} = 1 \\ y^{i-1}(m-1) + \alpha(y^{i-1}(m) - y^i(m-1)) & \text{m} > 1 \end{cases}$$

A.2 MGCEP formulation

If the input spectrum has a frequency warping as α_1 and the desired output frequency warping is α_2. Similarly, the input generalized cepstrum was estimated using the analysis (γ) parameter γ_1 and the desired output parameter is γ_2. The MGCEP recursion will give:

$$c^{(i)}_{\alpha_2,\gamma_1}(m) = \begin{cases} c_{\alpha_1,\gamma_1}(-i) + \alpha c^{(i-1)}_{\alpha_2,\gamma_1}(0) & \text{m} = 0 \\ (1-\alpha^2)c^{(i-1)}_{\alpha_2,\gamma_1}(0) + \alpha c^{(i-1)}_{\alpha_2,\gamma_1}(1) & \text{m} = 1 \\ c^{(i-1)}_{\alpha_2,\gamma_1}(m-1) + \alpha[c^{(i-1)}_{\alpha_2,\gamma_1}(m) - c^{(i)}_{\alpha_2,\gamma_1}(m-1)] & \text{m} = 2,3,\cdots,M_2 \end{cases}$$

$$\text{where, } i = -M_1,\cdots,-1,0 \qquad \text{and} \qquad \alpha = \frac{(\alpha_2-\alpha_1)}{(1-\alpha_1\alpha_2)} \tag{A.9}$$

$$K_{\alpha_2} = s^{-1}_{\gamma_1}(c^{(0)}_{\alpha_2,\gamma_1}(0)) \tag{A.10}$$

$$\hat{c}_{\alpha_2,\gamma_1}(m) = \frac{c^{(0)}_{\alpha_2,\gamma_1}(m)}{1+\gamma_1 c^{(0)}_{\alpha_2,\gamma_1}(0)} \quad \text{m} = 1,2,3,\cdots,M_2 \tag{A.11}$$

$$\hat{c}_{\alpha_2,\gamma_2}(m) = \hat{c}_{\alpha_2,\gamma_1}(m) + \sum_{k=1}^{m-1} \frac{k}{m}(\gamma_2 c_{\alpha_2,\gamma_1}(k)\hat{c}_{\alpha_2,\gamma_2}(m-k) - \gamma_1 c_{\alpha_2,\gamma_2}(k)\hat{c}_{\alpha_2,\gamma_1}(m-k)) \quad \text{(A.12)}$$

where, m = 1,2,3,$\cdots$,M_2

$$c_{\alpha_2,\gamma_2}(0) = s_{\gamma_2}(K_{\alpha_2}) \quad \text{(A.13)}$$

$$c_{\alpha_2,\gamma_2}(m) = \hat{c}_{\alpha_2,\gamma_2}(m)(1+\gamma_2 c_{\alpha_2,\gamma_2}(0)) \quad \text{m} = 1,2,3,\cdots,M_2 \quad \text{(A.14)}$$

The equations A.9 to A.14 represent the MGCEP recursion.This is derived from Oppenheim [1972] which converts a set of discrete representation into another set of discrete representation which preserves the convolution using the bilinear transform. This recursion can be represented into a matrix format.

The frequency transformation steps, Equation A.9 are derived in previous section. The Generalized logarithm is given by:

$$s_\gamma(\omega) = \begin{cases} \frac{(\omega^\gamma - 1)}{\gamma} & 0 < |\gamma| \leq 1 \\ \log\omega, & \gamma = 0 \end{cases}$$

The transfer function

$$H(z) = s_\gamma^{-1}(\sum_{m=0}^{\infty} c_\gamma(m) z^{-m}) \quad \text{(A.15)}$$

$$H(z) = \begin{cases} (1+\gamma\sum_{m=0}^{\infty} c_\gamma(m) z^{-m})^{\frac{1}{\gamma}} & 0 < |\gamma| \leq 1 \\ \exp\sum_{m=0}^{\infty} c_\gamma(m) z^{-m} & \gamma = 0 \end{cases}$$

When two different γ parameters are used, we can equate the spectrum and get the following results.

$$s_\gamma^{-1}(\sum_{m=0}^{\infty} c_\gamma(m) z^{-m}) = s_{\hat{\gamma}}^{-1}(\sum_{m=0}^{\infty} c_{\hat{\gamma}}(m) z^{-m}) \quad \text{(A.16)}$$

Differentiating this equation with respect to z^{-1}, and taking the inverse z-transform results in the γ part of the MGCEP recursion:

$$c_{\hat{\gamma}}(m) = c_\gamma(m) + \sum_{k=1}^{m-1} \frac{k}{m}[\hat{\gamma} c_\gamma(k) c_{\hat{\gamma}}(m-k) - \gamma c_{\hat{\gamma}}(k) c_\gamma(m-k)] \quad \text{(A.17)}$$

Combining Equation A.16 with the definition of inverse generalized log function,

$$(1+\gamma\sum_{m=0}^{\infty}c_{\gamma}(m)z^{-m})^{\frac{1}{\gamma}}=(1+\hat{\gamma}\sum_{m=0}^{\infty}c_{\hat{\gamma}}(m)z^{-m})^{\frac{1}{\hat{\gamma}}} \tag{A.18}$$

Taking log on both sides and differentiating with respect to z^{-1}.

$$\frac{d}{dz^{-1}}[\frac{1}{\gamma}\log(1+\gamma\sum_{m=0}^{\infty}c_{\gamma}(m)z^{-m})]=\frac{d}{dz^{-1}}[\frac{1}{\hat{\gamma}}\log(1+\hat{\gamma}\sum_{m=0}^{\infty}c_{\hat{\gamma}}(m)z^{-m})] \tag{A.19}$$

$$\frac{1}{(1+\gamma\sum_{m=0}^{\infty}c_{\gamma}(m)z^{-m})}\frac{d}{dz^{-1}}\sum_{m=0}^{\infty}c_{\gamma}(m)z^{-m}=\frac{1}{(1+\hat{\gamma}\sum_{m=0}^{\infty}c_{\hat{\gamma}}(m)z^{-m})}\frac{d}{dz^{-1}}\sum_{m=0}^{\infty}c_{\hat{\gamma}}(m)z^{-m} \tag{A.20}$$

$$\frac{\sum_{m=0}^{\infty}mc_{\gamma}(m)z^{-m+1}}{(1+\gamma\sum_{m=0}^{\infty}c_{\gamma}(m)z^{-m})}=\frac{\sum_{m=0}^{\infty}mc_{\hat{\gamma}}(m)z^{-m+1}}{(1+\hat{\gamma}\sum_{m=0}^{\infty}c_{\hat{\gamma}}(m)z^{-m})} \tag{A.21}$$

$$\sum_{m=0}^{\infty}mc_{\gamma}(m)z^{-m+1}(1+\hat{\gamma}\sum_{m=0}^{\infty}c_{\hat{\gamma}}(m)z^{-m})=\sum_{m=0}^{\infty}mc_{\hat{\gamma}}(m)z^{-m+1}(1+\gamma\sum_{m=0}^{\infty}c_{\gamma}(m)z^{-m}) \tag{A.22}$$

Equating the powers of z on both sides,

$$c_{\hat{\gamma}}(m)=c_{\gamma}(m)+\sum_{k=1}^{m-1}\frac{k}{m}[\hat{\gamma}c_{\gamma}(k)c_{\hat{\gamma}}(m-k)-\gamma c_{\hat{\gamma}}(k)c_{\gamma}(m-k)]$$

A.3 Spectral Criteria and its effects

The MGCEP analysis uses an optimization criterion similar to the UELS (unbiased estimation of log spectrum) to the spectral model. It can be shown that the minimization of this criterion is equivalent to the minimization of the mean square of the linear prediction error. As a result, this method can be viewed as a unified approach to speech spectral analysis, which includes several speech analysis methods. Although the method involves a non-linear minimization problem, it can easily be solved by an iterative algorithm [Tokuda et al., 1998, 1994b,a]. The convergence is quadratic and typically a few iterations are sufficient to obtain the

solution [Tokuda et al., 1994b]. The stability of the obtained model solution is also guaranteed. In the absence of this optimization criteria, the lower order MGCEP coefficients are just the truncated versions of higher order MGCEP coefficients.

B Cascade of All-pass transform based Warping

It can be shown that the cascade of two stages of the bilinear transform with warping parameters α_1 and α_2 is equivalent to applying a single stage of the bilinear transform with a combined warping factor

$$\alpha = \frac{(\alpha_1 + \alpha_2)}{(1 + \alpha_1\alpha_2)} \tag{B.1}$$

Let z be the complex variable in the original domain and s and u the complex variables after the two bilinear transformations respectively. The relationship connecting these domains are:

$$s^{-1} = \frac{(z^{-1} - \alpha_1)}{(1 - \alpha_1 z^{-1})} \tag{B.2}$$

$$u^{-1} = \frac{(s^{-1} - \alpha_2)}{(1 - \alpha_2 s^{-1})} \tag{B.3}$$

Substituting Equation B.2: in Equation B.3:

$$u^{-1} = \frac{\left(\frac{(z^{-1}-\alpha_1)}{(1-\alpha_1 z^{-1})} - \alpha_2\right)}{\left(1 - \alpha_2 \frac{(z^{-1}-\alpha_1)}{(1-\alpha_1 z^{-1})}\right)} = \frac{[z^{-1}(1 + \alpha_1\alpha_2) - (\alpha_1 + \alpha_2)]}{[(1 + \alpha_1\alpha_2) - (\alpha_1 + \alpha_2)z^{-1}]} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{B.4}$$

where α can be represented as:

$$\alpha = \frac{(\alpha_1 + \alpha_2)}{(1 + \alpha_1\alpha_2)} \tag{B.5}$$

C Deriving the quadratic differential formula

One of the standard matrix quadratic differential equations is given below:

$$\frac{\partial}{\partial \mathbf{z}}(\mathbf{z}^T \boldsymbol{A}\mathbf{z}) = \mathbf{z}^T(\boldsymbol{A}^T + \boldsymbol{A}) \tag{C.1}$$

since the result is a vector in row format, we could take a transpose to get the vector in column format which results in:

$$(\mathbf{z}^T(\boldsymbol{A}^T + \boldsymbol{A}))^T = \boldsymbol{A}\mathbf{z} + \boldsymbol{A}^T\mathbf{z} \tag{C.2}$$

Substituting variable $z = y - x$,

$$\frac{\partial \mathbf{z}}{\partial \boldsymbol{x}} = -1$$

$$\frac{\partial}{\partial \mathbf{z}}(\mathbf{z}^T \boldsymbol{A}\mathbf{z}) = \Big[\boldsymbol{A}(\mathbf{y} - \boldsymbol{x}) + \boldsymbol{A}^T(\mathbf{y} - \boldsymbol{x})\Big] \times -1$$

which results in the equation.

$$\frac{\partial}{\partial x}(\boldsymbol{y} - \boldsymbol{x})^T \boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x}) = -\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x}) - \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{x}) \tag{C.3}$$

D Summary of databases

There are different databases that were evaluated in this work. Following is a short summary of these databases.

D.1 WSJ0

Wall street journal (WSJ) is an American English database that was recorded to support research on large-vocabulary Continuous Speech Recognition (CSR) systems. WSJ0 was the first database of WSJ (WSJ0 and WSJ1) series recorded in 1991.

The corpora consist primarily of read speech with texts drawn from a machine-readable corpus of Wall Street Journal news text. The texts to be read were selected to fall within either a 5,000-word or a 20,000-word subset of the WSJ text corpus. Some spontaneous dictation was included in addition to the read speech. The dictation portion was collected using journalists who dictated hypothetical news articles.

Two microphones were used in this database: a close-talking Sennheiser HMD414 and a secondary microphone, which may vary. The corpora were thus offered in three configurations: the speech from the Sennheiser, the speech from the other microphone and the speech from both; all three sets include all transcriptions, tests, documentation, etc.

D.1.1 SI-84 set

The speech was recorded at 16kHz and 39-dimensional MGCEP features were extracted from this data for TTS. This training subset consists of 83 speakers and referred to as speaker independent (SI-84) system. This includes balanced male and female speakers. The speaker adaptive (SAT) models were generated using this training data. For the unified models in ASR, the same data was used to extract 13-dimensional MGCEP features and SAT models were trained. Unilex phoneset was used to generate the full-context labels required for TTS.

D.1.2 Evaluation sets

In early 1993, the ARPA CSR Corpus Coordinating Committee (CCCC) designed a "Hub and Spoke" test paradigm. The resulting suite of tests contained two general "hub" tests to assess the performance of large (5K vocabulary) and very large (64K vocabulary) speaker-independent continuous speech recognition systems and nine "spoke" tests to assess the performance of systems designed to address specific areas of research in continuous speech recognition. The evaluations are named the Nov93 evaluations.

The goal of the "hub" task is to improve basic SI performance on clean data. This set includes 10 speakers with 20 utterances each The 5K-word vocabulary as the read WSJ data was recorded using the Sennheiser microphone. The specific hub condition tested in this work is the "P0" option. The "P0" set can use any grammar or acoustic training, and includes session boundaries and utterance order given as side information.

The other evaluation set used in this work is the "spoke" task with incremental speaker adaptation. The goal is to evaluate an incremental speaker adaptation algorithm with 40 adaptation sentences from four speakers. There are 100 utterances from these speakers for evaluations. The set has 5K vocabulary read WSJ data was recorded using the Sennheiser microphone. The set used for evaluations in this work was the "C2" and "C3" options which are the incremental supervised adaptation and the rapid enrollment speaker adaptation tasks respectively.

D.1.3 Gender dependent models

The training data for WSJ0 has a lot more speakers than 83 speakers. The data has approximately equal number of males and females in the training set. Gender dependent male models were generated using 59 male speakers from the training set. Gender dependent female models were trained using the data from 60 female speakers of the same set. The data was recorded at 16kHz sampling rate and 39-dimensional MGCEP features were used to generate the gender dependent models. Unilex phoneset based full-context labels were used to generate the TTS models.

D.1.4 Test speakers

There are more test speakers in this database. There are speakers classified as for development test and also speakers for evaluation tests. 30 speakers from the development test set were used in the evaluations presented in chapter 4.

D.2 WSJCAM0

The Wall Street Journal recorded at the University of CAMbridge (WSJCAM0) speech database is the UK English equivalent of a subset of the US American English WSJ0 database. It consists of speaker-independent (SI) read material, split into training, development test and evaluation test sets. There are 90 utterances from each of 92 speakers that are designated as training material for speech recognition algorithms. Recordings were made from two microphones: a far-field desk microphone and a head-mounted close-talking microphone. The recorded sentences were same as the WSJ American text corpus.

A modified pronunciation dictionary was constructed that covered UK pronunciations for some US-specific words. The training sentences were taken from the WSJ0 training sub corpus of about 10,000 sentences. 48 test speakers recorded 80 sentences (with 40 sentences from 5K word vocabulary and 40 sentences from 64K word corpus). Similar to the tests with WSJ0, 30 test speakers were selected from the test sets. Data were sampled at 16kHz and 39-dimensional MGCEP features were extracted to build the TTS models.

D.3 Databases recorded at CSTR

Several databases were collected at the "Centre for Speech Technology Research" (CSTR), Edinburgh. All the databases were recorded at the anechoic recording studio at CSTR.

D.3.1 Gender dependent database

Native UK English speakers recorded data at a sampling rate of 96kHz. The training set is made of 31 male and 29 female speakers. This data was used to generate gender dependent male and female models respectively. The data was down sampled to 48kHz and 59-dimensional feature vectors and full-context labels using Combilex RPX phoneset were used to build the TTS models. The test set includes 31 male speakers and 30 female speakers. 100 sentences are available for adaptation and testing for each test speaker.

D.3.2 Child speech

Two child speakers were recorded at this same recording studio. The children were asked to read fairy tales and the data was recorded at 96kHz (and down-sampled to 48kHz). The first child has only four aligned labeled adaptation utterances. The labels for synthesis are the three fairy tales of 67 sentences in total. A fairy tale consisting of 100 utterances were recorded by the second child at 96kHz sampling rate. These sentences were also down-sampled to 48kHz. This data was used for objective evaluations.

D.3.3 EMIME bilingual database

In the context of EMIME project, Bilingual databases were recorded for German/English, Finnish/English and Mandarin/English language pairs at CSTR. This work uses the Mandarin/English Bilingual database. It includes the recordings of seven female and seven male speakers of Mandarin. Two different microphones were used, a close-talking DPA 4035 mounted on the subjects headphones and a Sennheiser MKH 800 p48 microphone placed about 10cm from subject using an omnidirectional pattern. The speech was sampled at 96kHz 24bit depth and stored directly to a computer. These recordings were subsequently down sampled, using Pro-Tools to 22 kHz 16bit and released. The data was further down sampled to 16kHz and 39-dimensional MGCEP features were extracted for using the samples as adaptation data in this work. The news sentences for English were taken from the WSJ1 corpus comprising 40 enrollment sentences and 60 test set sentences. The Mandarin news sentences were selected from the Speecon corpus.

D.4 Speecon Mandarin

The Mandarin Chinese Speecon database comprises of the recordings of 550 adult Chinese speakers (276 males, 274 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place). Each of the four speech channels is recorded at 16 kHz, 16 bit, uncompressed unsigned integers in Intel format (lo-hi byte order). To each signal file corresponds an ASCII SAM label file which contains the relevant descriptive information. 39-dimensional MGCEP features were used to generate Mandarin average voice model.

D.5 Aurora4

The Aurora4 database is the noisy version of the WSJ0 database. The noise signals have been artificially added to the fairly clean data. There exist versions at sampling rates of 8 and 16 kHz. Six noise conditions collected from street traffic, train stations, cars, babble, restaurants and airports were digitally added to the speech data to simulate degradations in the signal-to-noise ratio of the channel. Both microphone conditions (Sennheiser and Second Microphone) contained in the WSJ0 corpus were modified.

D.6 Blizzard database

The Blizzard challenge 2010 database consists of two male RP English speakers (RJS and Roger). The RJS corpus was provided by Phonetic Arts, and the Roger corpus was from the University of Edinburgh. The speaker dependent models were build sung the 4014 utterances from RJS and Roger was used as the test speaker. The data is RIFF with little-endian (Microsoft PCM) format. The waveforms are 16 bit with a sampling rate of 48kHz. Full-context labels were

generated using the Unilex RPX phoneset. 59 dimensional MGCEP features were extracted from these data. The objective evaluations performed for Roger was based on the 'ES1' task of Blizzard challenge 2010. The task was to build voices from the first 100 utterances of the Roger database using voice conversion, speaker adaptation or similar techniques. Approximately 468 sentence labels were offered by the Blizzard challenge 2010 for different subjective evaluations including broadcast news, read news, novel and semantically unpredictable sentences. These were used for synthesizing the sentences for the subjective evaluations presented in this thesis.

D.7 EMIME noisy speech

Database was collected in the framework of EMIME project at a reputed speech conference (Interspeech 2008). Some of the participants in the conference were asked to read a set of pre-selected text (mainly newspaper articles). 67 participants participated and included both native and non-native English speakers. Since the data was recorded in the noisy background of a conference setting, the speech has babble noise. Approximately 100 sentences were recorded by each speaker as adaptation and test data. There were fewer female speakers compared to male speakers. The data was recorded in RIFF file format at 44100Hz sampling rate. The waveforms were up sampled to 48kHz for the evaluations performed in this thesis. Full-context labels were generated using the Combilex RPX phoneset. Forty five speakers (including six females) speaking the same text and with adequate adaptation data were selected for the evaluations.

Bibliography

A. Acero. *Acoustical and environmental robustness in automatic speech recognition*. Springer, US, 1993.

A. Acero and R. M. Stern. Robust speech recognition by normalization of the acoustic space. In *Proc. of ICASSP, Pages = 893-896, Address = Toronto, Canada, year = 1991*.

Mohamed Afify and Olivier Siohan. Constrained maximum likelihood linear regression for speaker adaptation. In *Proc. of Interspeech*, pages 861–864, 2000.

P. T. Akhil, S. P. Rath, S. Umesh, and D. R. Sanand. A computationally efficient approach to warp factor estimation in VTLN using EM algorithm and sufficient statistics. In *Proc. of Interspeech*, pages 1713–1716, Brisbane, Australia, 2008.

Alan W. Black, Heiga Zen, and Keiichi Tokuda. Statistical parametric speech synthesis. In *Proc. of ICASSP*, pages 1229–1232, Hawaii, USA, 2007.

C. Breslin, K.K. Chin, M.J.F. Gales, K. Knill, and H. Xu. Prior information for rapid speaker adaptation. In *Proc. of Interspeech*, pages 1644–1647, Japan, 2010.

W. Chou. Maximum a posterior linear regression with elliptically symmetric matrix variate priors. In *Proc. of Eur. Conf. Speech Communication Technology*, Budapest, Hungary, 1999.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38, 1977.

V. Digalakis, D. Rtischev, L. Neumeyer, and Edics Sa. Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 3: 357–366, 1995.

J. Dines, L. Saheer, and H. Liang. Speech recognition with synthesis models by marginalising over decision tree leaves. In *Proc. of Interspeech*, pages 1395–1398, September 2009a.

J. Dines, J. Yamagashi, and S. King. Measuring the gap between HMM-based ASR and TTS. In *Proc. of Interspeech*, pages 1391–1394, September 2009b.

John Dines, Junichi Yamagishi, and Simon King. Measuring the gap between hmm-based asr and tts. *IEEE Journal of Selected Topics in Signal Processing*, 4(6):1046–1058, December 2010.

E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. of ICASSP*, pages 346–348, Washington, DC, USA, 1996.

T. Emori and K. Shinoda. Rapid vocal tract length normalization using maximum likelihood estimation. In *Proc. of Eurospeech*, pages 1649–1652, 2001.

Arlo Faria and David Gelbart. Efficient pitch-based estimation of VTLN warp factors. In *Proc. of Interspeech*, pages 213–216, September 2005.

F. Flego and M. J. F. Gales. Incremental predictive and adaptive noise compensation. In *Proc. of ICASSP*, pages 3837–3840, Washington, DC, USA, 2009.

Toshiaki Fukada and Yoshinori Sagisaka. Speaker normalized acoustic modeling based on 3-d viterbi decoding. In *Proc. of ICASSP*, pages 437–440, Seattle, USA, 1998.

M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12 (2):75–98, 1998.

M.J.F. Gales and P.C. Woodland. Mean and variance adaptation within the mllr framework. *Computer Speech & Language*, 10:249–264, 1996.

Giulia Garau. *Speaker Normalization for Large Vocabulary Multiparty Conversational Speech Recognition*. PhD thesis, University of Edinburgh, 2008.

Philip N. Garner. Speech signal processing. Idiap-Com Idiap-Internal-Com-02-2010, Idiap, 3 2010.

Philip N. Garner and Wendy J. Holmes. On the robust incorporation of formant features into hidden Markov models for automatic speech recognition. In *Proc. of ICASSP*, volume 1, pages 1–4, 1998.

Matthew Gibson and William Byrne. Unsupervised intralingual and cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction. *IEEE Transactions on Audio, Speech, and Language Processing (In print)*, 2010.

M. Hirohata, T. Masuko, and T. Kobayashi. A study on average voice model training using vocal tract length normalization. *IEICE Technical Report,*, 103 (27):69–74, 2003. In Japanese.

Reima Karhila, Rama Sanand Doddipatla, Mikko Kurimo, and Peter Smit. Creating synthetic voices for children by adapting adult average voice using stacked transformations and VTLN. In *Proc. of ICASSP*, pages 4501–4504, Kyoto, Japan, March 2012.

Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3-4):187–207, 1999.

D. Y. Kim, S. Umesh, M. J. F. Gales, T. Hain, and P. C. Woodland. Using VTLN for broadcast news transcription. In *Proc. of ICSLP*, pages 1953–1956, South Korea, 2004.

Simon King, Keiichi Tokuda, Heiga Zen, and Junichi Yamagishi. Unsupervised adaptation for hmm-based speech synthesis. In *Proceedings of Interspeech*, pages 1869–1872, Brisbane, Australia, September 2008.

Li Lee and Richard Rose. A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing,*, 6:49–60, 1998.

Li Lee and Richard C. Rose. Speaker normalization using efficient frequency warping procedures. In *Proc. of ICASSP*, pages 353–356, Washington, DC, USA, 1996.

C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9:171–186, 1995.

Chris J Leggetter. *IMPROVED ACOUSTIC MODELLING FOR HMMS USING LINEAR TRANSFORMATIONS*. PhD thesis, University of Cambridge, 1995.

Hui Liang and John Dines. An analysis of language mismatch in HMM state mapping-based cross-lingual speaker adaptation. In *Proc. of Interspeech*, pages 622–625, September 2010.

Hui Liang, John Dines, and Lakshmi Saheer. A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis. In *Proc. of ICASSP*, pages 4598–4601, March 2010.

Louis A. Liporace. Maximum-likelihood estimation for multivariate observations of markov sources. *IEEE Transactions on Information Theory*, 28(5):729–734, 1982.

Jonas Lööf, Christian Gollan, and Hermann Ney. Speaker adaptive training using shift-mllr. In *Proceedings of Interspeech*, pages 1701–1704, Brisbane, Australia, September 2008.

J. McDonough, G. Zavaliagkos, and H. Gish. An approach to speaker adaptation based on analytic functions. In *Proceedings of ICASSP*, pages 721–724, Atlanta, USA, May 1996.

J. W. McDonough. *Speaker Compensation with All-Pass Transforms*. PhD thesis, John Hopkins University, 2000.

A. Miguel, E. Lleida, R. L. Buera, and A. Ortega. Augemented state space acoustic decoding for modeling local variability in speech. In *Proc. of Interspeech*, Lisbon, Portugal, 2005.

A. Miguel, E. Lleida, R. L. Buera, A. Ortega, and O. Saz. Local transformation models for speech recognition. In *Proc. of Interspeech*, Pittsburg, USA, 2006.

A. Miguel, E. Lleida, R. Rose, L. Buera, O. Saz, and A. Ortega. Capturing local variability for speaker normalization in speech recognition. *Transactions on Audio, Speech and Language Processing*, 16(3):578–593, March 2008.

Daisuke Miyamoto, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. Acoustic compensation methods for body transmitted speech conversion. In *Proc. of ICASSP*, pages 3901–3904, April 2009.

Sirko Molau, S. Kanthak, and Hermann Ney. Efficient vocal tract normalization in ASR. In *Proc. of ESSV*, Cottbus, Germany, 2000.

D.H. Oppenheim, A.V. Johnson. Discrete representation of signals. *Proc. of IEEE*, 60:681–691, 1972.

Keiichiro Oura, Keiichi Tokuda, Junichi Yamagishi, Simon King, and Mirjam Wester. Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis. In *Proc. of ICASSP*, pages 4594–4597, March 2010.

Sankaran Panchapagesan and Abeer Alwan. Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC. *Computer Speech & Language*, 23(1):42–64, 2009.

M. Pitz and H. Ney. Vocal tract normalization equals linear transformation in cepstral space. *IEEE Transactions on Speech and Audio Processing*, 13:930–944, 2005.

Michael Pitz. *Investigations on Linear Transformations for Speaker Adaptation and Normalization.* PhD thesis, RWTH Aachen University, 2005.

William Press, Saul Teukolsky, William Vetterling, and Brian Flannery. *Numerical Recipes in C.* Cambridge University Press, 1992.

D. Pye and P. C. Woodland. Experiments in speaker normalisation and adaptation for large vocabulary speech recognition. In *Proc. of ICASSP*, pages 1047–1050, 1997.

S. P. Rath and S. Umesh. Acoustic class specific VTLN-warping using regression class trees. In *Proc. of Interspeech*, pages 556–559, Brighton, UK, 2009.

S. P. Rath, S. Umesh, and A. K. Sarkar. Using VTLN matrices for rapid and computationally-efficient speaker adaptation with robustness to first-pass transcription errors. In *Proc. of Interspeech*, pages 572–575, Brighton, UK, 2009.

R. Rose, A. Keyvani, and A. Miguel. On the interaction between speaker normalization, environment compensation, and discriminant feature space transformations. In *Proc. of ICASSP*, Toulouse, France, 2006.

Lakshmi Saheer, John Dines, Philip N. Garner, and Hui Liang. Implementation of VTLN for statistical speech synthesis. In *Proc. of the 7th ISCA Speech Synthesis Workshop*, pages 224–229, Kyoto, Japan, September 2010a.

Lakshmi Saheer, Philip N. Garner, and John Dines. Study of Jacobian normalization for VTLN. *Idiap-RR-25-2010*, 2010b.

Lakshmi Saheer, Philip N. Garner, John Dines, and Hui Liang. VTLN adaptation for statistical speech synthesis. In *Proc. of ICASSP*, pages 4838–4841, March 2010c.

Lakshmi Saheer, John Dines, and Philip N. Garner. Vocal tract length normalization for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 20(7):2134–2148, 2012a.

Lakshmi Saheer, Hui Liang, John Dines, and Philip N. Garner. Vtln-based rapid cross-lingual adaptation for statistical parametric speech synthesis. *Idiap-RR-12-2012*, 2012b.

Lakshmi Saheer, Junichi Yamagishi, Philip N. Garner, and John Dines. Combining vocal tract length normalization with hierarchial linear transformations. In *Proc. of ICASSP*, pages 4493–4496, Kyoto, Japan, March 2012c. IEEE SPS.

D. R. Sanand, S. P. Rath, and S. Umesh. A study on the influence of covariance adaptation on Jacobian compensation in vocal tract length normalization. In *Proc. of Interspeech*, pages 584–587, Brighton, UK, 2009.

D. Rama Sanand and Srinivasan Umesh. VTLN using analytically determined linear-transformation on conventional mfcc. *IEEE Transactions on Audio, Speech & Language Processing*, 20(5):1573–1584, 2012.

Ananth Sankar and Chin-Hui Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(3):190–202, May 1996.

Oscar Saz, Antonio Miguel, Eduardo Lleida, Alfonso Ortega, and Luis Buera. Study of time and frequency variability in pathological speech and error reduction methods for automatic speech recognition. In *Proc. of Interspeech*, pages 993–996, Pittsburgh, USA, 2006.

K. Shinoda and C. Lee. A structural Bayes approach to speaker adaptation. *IEEE Transactions on Speech Audio Processing*, 9:276–287, March 2001.

O. Shiohan, T. Myrvoll, and C. Lee. Structural maximum a posteriori linear regression for fast HMM adaptation. *Computer, Speech and Language*, 16(3):5–24, January 2002.

David Sündermann. *Text-Independent Voice Converision*. PhD thesis, Bundeswehr University Munich, Munich, Germany, 2008.

Keiichi Tokuda, Takao Kobayashi, and Santoshi Imai. Recursive calculation of mel-cepstrum from lp coefficients. *Technical Report*, pages 1–7, 1994a. URL http://www.sp.nitech.ac.jp/~tokuda/tips/.

Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai. Mel-generalized cepstral analysis – A unified approach to speech spectral estimation. In *Proc. of ICSLP*, volume 3, pages 1043–1046, September 1994b.

Keiichi Tokuda, Takao Kobayashi, and Santoshi Imai. Recursion formula for calculation of mel generalized cepstrum coefficients (in Japanese). *Trans. IEICE,*, J71-A:128–131, 1998.

L. F. Uebel and P. C. Woodland. An investigation into vocal tract length normalisation. In *Proc. of the European Conference on Speech Communication and Technology*, pages 2527–2530, 1999.

S. Umesh, A. Zolnay, and H. Ney. Implementing frequency warping and VTLN through linear transformation of conventional MFCC. In *Proc. of Interspeech*, pages 269–271, Lisbon, Portugal, 2005.

Balakrishnan Varadarajan, Daniel Povey, and Selina M. Chu. Quick fMLLR for speaker adaptation in speech recognition. In *Proc. of ICASSP*, pages 4297–4300. IEEE, 2008.

Hisashi Wakita. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(2):183–192, April 1977.

L. Welling, S. Kanthak, and H. Ney. Improved methods for vocal tract normalization. In *Proc. of ICASSP*, pages 761–764, 1999.

Mirjam Wester and Hui Liang. The EMIME Mandarin bilingual database. Technical Report EDI-INF-RR1396, University of Edinburgh, U.K., February 2011a.

Mirjam Wester and Hui Liang. Cross-lingual speaker discrimination using natural and synthetic speech. In *Proc. of Interspeech*, pages 2481–2484, August 2011b.

Yi-Jian Wu, Yoshihiko Nankaku, and Keiichi Tokuda. State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. In *Proc. of Interspeech*, pages 528–531, September 2009.

Junichi Yamagashi, Heiga Zen, Tomoki Toda, and Keichi Tokuda. Speaker independent HMM based speech synthesis system - HTS-2007 system for blizzard challenge 2007. In *Proc. of Sixth ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007.

Junichi Yamagishi, Takao Kobayashi, Yuji Nakano, Katsumi Ogata, and Juri Isogai. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1): 66–83, January 2009a.

Junichi Yamagishi, Takashi Nose, Heiga Zen, Zhen-Hua Ling, Tomoki Toda, Keiichi Tokuda, Simon King, and Steve Renals. Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6):1208–1230, August 2009b.

Junichi Yamagishi, Oliver Watts, Simon King, and Bela Usabaev. Roles of the average voice in speaker-adaptive HMM-based speech synthesis. In *Proc. of Interspeech*, pages 418–421, September 2010.

H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Hidden semi-Markov model based speech synthesis. In *Proc. of ICSLP*, pages 1397–1400, Korea, 2004.

H. Zen, K. Tokuda, and A. W. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, November 2009.

Xin Zhuang, Yao Qian, Frank K. Soong, Yi-Jian Wu, and Bo Zhang. Formant-based frequency warping for improving speaker adaptation in hmm tts. In *Proc. of Interspeech*, pages 817–820, Makuhari, Japan, 2010.

Lakshmi B. Saheer

Avenue des Alpes 54
1820 Montreux
+41-786060498
+41-27-7217789
lsaheer@idiap.ch

Profile

Goals — Develop and improve a comprehensive understanding of speech technology. The current focus of my research is to make a lasting contribution in building the "EMIME" personalized speech-to-speech translation system.

Experience — Currently pursuing my doctoral studies jointly at Idiap research institute and Ecole Polytechnique Fédérale de Lausanne. I come from a Computer Science background with 6 years of research experience focussed on different speech technologies and 3 years of software engineering experience with international companies.

Skills — Good Knowledge of speech technologies such as HMM-based synthesis and recognition. My skillset includes most modern speech and programming toolkits.

Education

2008–2012 (expected) — **PhD**, *Ecole Polytechnique Fédérale de Lausanne (EPFL)*, Switzerland.

2004–2007 — **M.S in Computer Science**, *Indian Institute of Technology Madras*, India, *8.4/10*.

1998–2002 — **B.Tech in Computer Science**, *Cochin University of Science & Technology*, India, *84%*.

PhD thesis

title — *Effective Multilingual Interaction in Mobile Environments (EMIME)*

supervisors — Prof. Hervé Bourlard, Philip Garner & Dr. John Dines

description — The EMIME project is focused on personalized speech-to-speech translation, such that the speech in one language is translated and synthesized in the same user's voice in another language. The recent developments in speech synthesis using hidden Markov models, which is the same technology used for automatic speech recognition is used to unify the process of recognition and synthesis. Using a common statistical modeling framework for automatic speech recognition and speech synthesis will enable the use of common techniques for adaptation and multilinguality. ASR systems use features that are insensitive to sources of variation such as speaker identity. While, the Text to Speech (TTS) system tries to capture/model the personal characteristics of a speaker (like the pitch information) so that the synthesized voice is similar to the speaker. My doctoral research is focused on finding common features for the recognizer and synthesizer so that the speaker adaptation and other related techniques can be applied to this multilingual translation system easily.

Master thesis

title — *Syllable based Speech Recognition for Indian Languages*

supervisors — Prof. Hema A. Murthy & Dr. C. S. Ramalingam

description Building large vocabulary continuous speech recognition systems has been an important area of research for more than a decade. The current day challenge in building such a system is the cost of obtaining necessary annotated transcribed train data. With this task being both time consuming and laborious, this work analyzes a novel technique for building a syllable based continuous speech recognizer when only unannotated transcribed train data is available. Group delay based segmentation is modified to give accurate syllable boundaries even for fricatives, long silences and semi-vowels. Tamil (a regional Indian language) has been studied to collect generic rules to syllabify text. This generic algorithm, though developed for Tamil, can be used for any syllable based language. The syllabified text can then be used to automatically annotate speech into syllable units. Isolated style syllable models are built using Multiple Frame Size (MFS) and Multiple Frame Rate (MFR) based feature extraction techniques, taking into account the spectral changes in the speech data. Experiments performed on Indian languages like Tamil and Hindi show that the recognition performance is comparable to recognizers built using manually segmented train data. Word error rate of 25% is achieved for Tamil by applying bigram language models on the recognizer output. These experiments suggest that system development cost can be reduced with minimum manual effort, if sentence level transcription of the speech data is available.

Experience

2008–Present **Research Assistant**, *Idiap Research Institute*, Martigny, Switzerland.
Involved in the EMIME personalized speech to speech translation project in the unified framework of hidden Markov models. Working on unifying the adaptation techniques for recognition and synthesis. Implementing the rapid speaker adaptation techniques for synthesis. Vocal tract length normalization is a rapid adaptation technique commonly used in ASR and involves a lot of challenges when applying to TTS (like the higher order features and Jacobian Normalization). Current work has successfully implemented VTLN in the framework of statistical speech synthesis and speaker characteristics are reproduced in the synthesized speech with very little adaptation data (as small as a single adaptation utterance).

2008–2008 **Software Engineer**, *Sony Ericsson AB*, Lund, Sweden.
Building Mobile Applications for Sony Ericsson mobile phones. Worked with the Network module specially the GAN (Generic Access Network) Application. The applications are built on Ericsson Mobile Platforms (EMP). Handled UI and Service layer applications. Familiar with the Sony Ericsson proprietary tools for Flashing, Target Debugging and Simulation of Applications for Mobile Phones. Technologies: C, C++.

2007–2007 **Research Analyst**, *V-Enable Software Pvt. Ltd.*, India.
Improving Speech Recognition Performance (Nuance Recognizer). Various signal processing techniques were implemented for improving the performance of Nuance based commercial speech recognizer for mobile phone. Speech endpoint detection, inter speech silence removal, SNR detection and noise removal techniques were tried to check the recognition performance improvement for noisy input speech. Technologies: Nuance recognizer, C, C++, Perl, shell scripting

2004–2007 **Project Associate**, *TeNet group, Indian Institute of Technology*, Madras.
Continuous Speech Recognizers (CSR) for Tamil. CSR was implemented as a part of Multimodal Interface to the computer for Indian languages. Initially, SPHINX-3 was used to build a CSR system using broadcast news and telephone data. Then, a syllable based CSR was built using HMM Toolkit (HTK) by generating automatically annotated train data using acoustic and text segmentation algorithms developed. Language transliteration tool, text syllabification tool, language model statistics generator tools and search algorithms have also been implemented as a part of this project. Technologies: SPHINX, HTK, C, C++, Perl, shell scripting

2002–2004 **System Engineer**, *Siemens Information Systems Ltd.*, India.
Common Electronic Patient Record (Soarian) Soarian is an IT-solution for healthcare organizations like hospitals and primary care services all over the world. Clinical modules can be plugged in and out of the system. This can be used in small to large hospitals and even in a network of geographically separated hospitals sharing a common patient database. There are a wide range of user interfaces: desktop, web-interface, laptop and palmtop computers. The 3-tier architecture and network distributed databases are the key design features of this system. Technologies: COM, C++, Visual Studio, SQL, XML, HTML

Languages

English	**Fluent**	*Medium of Education since school*
Malayalam	**Native**	
Hindi, Tamil	**Fluent**	
French	**Basic**	

Computer skills

Programming:	C, C++, Visual Basic, MATLAB, Perl, shell scripting
Speech toolkits:	SPHINX, HTK, HTS, NUANCE, SRILM (Language modeling)
SP tools:	Edinburgh Speech Tools, SPTK
Other tools:	gdb, ddd, Rational Rose, CharmNT, Rational ClearCase

Research Interests

- HMM Speech Recognition
- Statistical Speech Synthesis
- Feature Adaptation/Transformation
- Common Feature representations for Speech Recognition and Synthesis
- Pattern Recognition/Classification

Professional activities and Awards

- Google Anitaborg Scholarship recipient 2011.
- Contribution to the open source speech synthesis toolkit HTS (VTLN Extension).
- Reviewer of IEEE Transactions on Audio, Speech and Language Processing.
- Student Member of IEEE & IEEE SPS.
- Consolation prize for SWAT THE BUG contest for Shaastra-2006, IIT Madras, India.
- 3rd rank holder at University level for Bachelors (2002).
- Best outgoing student in school (1996).

Personal Details

Nationality	Indian
Marital Status	Married
Maiden Name	Achuthankutty

Referees

- Prof. Hervé Bourlard,
  Director, Idiap Research Institute,
  Professor, EPFL, Switzerland.
  Herve.Bourlard@idiap.ch

- Philip N. Garner,
  Senior Researcher, Idiap Research Institute,
  Switzerland. phil.garner@idiap.ch

- Dr. John Dines,
  Senior Researcher, Idiap Research Institute,
  Switzerland. john.dines@idiap.ch

Publications

Lakshmi Saheer, Junichi Yamagishi, John Dines, and Philip N. Garner. Combining vocal tract length normalization with hierarchial linear transformations. In *Proceedings of ICASSP*, pages 4493–4496, Kyoto, Japan, 2012.

Lakshmi Saheer, Hui Liang, John Dines, and Philip N. Garner. VTLN-based rapid cross-lingual adaptation for statistical parametric speech synthesis. *Idiap Research Report*, 2011.

Lakshmi Saheer, John Dines, Philip N. Garner, and Hui Liang. Implementation of VTLN for statistical speech synthesis. In *Proceedings of 7th ISCA Speech Synthesis Workshop (SSW7)*, pages 224–229, Kyoto, Japan, 2010.

Lakshmi Saheer, Philip N. Garner, John Dines, and Hui Liang. VTLN adaptation for statistical speech synthesis. In *Proceedings of ICASSP*, pages 4838–4841, DALLAS, USA, 2010.

Mirjam Wester, John Dines, Matthew Gibson, Hui Liang, Yi-Jian Wu, and **Lakshmi Saheer** et.al. Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *Proceedings of 7th ISCA Speech Synthesis Workshop (SSW7)*, pages 192–197, Kyoto, Japan, 2010.

Mikko Kurimo, William Byrne, John Dines, Philip N. Garner, and **Lakshmi Saheer** et.al. Personalising speech-to-speech translation in the EMIME project. In *Proceedings of the ACL 2010 System Demonstrations*, pages 48 – 53, Uppsala, Sweden, 2010.

Lakshmi Saheer, Philip N. Garner, and John Dines. Study of Jacobian normalization for VTLN. *Idiap Research Report*, 2010.

Hui Liang, John Dines, and **Lakshmi Saheer**. A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis. In *Proceedings of ICASSP*, pages 4598–4601, DALLAS, USA, 2010.

John Dines, **Lakshmi Saheer**, and Hui Liang. Speech recognition with synthesis models by marginalising over decision tree leaves. In *Proceedings of Interspeech*, pages 1395–1398, UK, 2009.

Lakshmi A. and Hema A. Murthy. A new approach to continuous speech recognition in Indian languages. In *Proceedings of NCC*, pages 277–281, India, 2008.

Lakshmi A. and Hema A. Murthy. A syllable based continuous speech recognizer for Tamil. In *Proceedings of ICSLP*, page 1878–1881, Pittsburgh, USA, 2006.

References to Journals

Lakshmi Saheer, Junichi Yamagishi, Philip N. Garner, and John Dines. Combining vocal tract length normalization with hierarchial linear transformations. *IEEE Transactions on Audio, Speech and Language Processing (to be submitted)*, 2012.

Lakshmi Saheer, John Dines, and Philip N. Garner. Vocal tract length normalization for

statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 20(7):2134–2148, 2012.

John Dines, Hui Liang, and **Lakshmi Saheer** et.al. Personalising speech-to-speech translation: Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis. *Computer Speech & Language Processing (Accepted for publication)*, 2011.

Hervé Bourlard, John Dines, Mathew Magimai-Doss, Philip N. Garner, David Imseng, Petr Motlicek, Hui Liang, **Lakshmi Saheer**, and Fabio Valente. Current trends in multilingual speech processing. *Invited paper in SADHANA*, 2010.

G. L. Sarada, **Lakshmi A.**, Nagarajan T., and Hema A. Murthy. Automatic transcription of continuous speech into syllable-like units for Indian languages. *Proc. of SADHANA, Journal on Academy Proceedings in Engineering Sciences*, 34:221–233, 2009.